Analysis and Optimization of Global Interconnects for Many-Core Architectures

A Thesis
Presented to
The Academic Faculty

by

Anant Balakrishnan

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2010

Analysis and Optimization of Global Interconnects for Many-Core Architectures

Approved by:

Dr. Azad Naeemi, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Jeffrey A. Davis
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: December 2010

To Amma, Appa, Paati, Abhi and Dallu

Acknowledgements

I would like to sincerely thank my advisor, Professor Azad Naeemi, for giving me an opportunity to work and learn under him. He is a wonderful mentor, excellent teacher and a tremendous source of inspiration. This thesis was possible because of the guidance and support I received from him at all times. It has been a privilege to have worked under his supervision.

I am grateful to Professors Sudhakar Yalamanchili and Jeff Davis, for their valuable time and for agreeing to be a part of my thesis committee. I would also like to thank Professors Saibal Mukhopadhyay and Gabriel Alfonso Rincón-Mora, for teaching me the art of circuit design.

I would like to extend special thanks to all my colleagues at Nanoelectronics Research Lab: Shaloo Rakheja, Omer Jamal, Ahmet Ceyhan and Vachan Kumar, for their deep insights in electronics and for making my workplace exciting.

I am extremely thankful to my parents, Meena and Balakrishnan, grandma, Lakshmy, brother, Abhishek and the love of my life, Sharmila, for their unconditional trust, support and encouragement throughout the course of this thesis. Many, many thanks to one and all.

Table of Contents

List of Figures

Summary

The objective of this thesis is to develop circuit-aware interconnect technology optimization for network-on-chip based many-core architectures. The dimensions of global interconnects in many-core chips are optimized for maximum bandwidth density and minimum delay taking into account network-on-chip router latency and size effects of copper. The optimal dimensions thus obtained are used to characterize different network-on-chip topologies based on wiring area utilization, maximum core-to-core channel width, aggregate chip bandwidth and worse case latency. Finally, the advantages of many-core many-tier chips are evaluated for different network-on-chip topologies. Area occupied by a router within a core is shown to be the bottleneck to achieve higher performance in network-on-chip based architectures.

Chapter 1

Introduction

1.1 Moore's Law

Processors have grown at an exponential rate in terms of number of transistors per die and performance since 1960's. This trend as shown in Figure 1.1 is known as Moore's Law, predicted in 1965 by Gordon Moore [1]. The driving factor for the semiconductor industry to follow this trend has been the development of smaller and faster transistors every technology year.

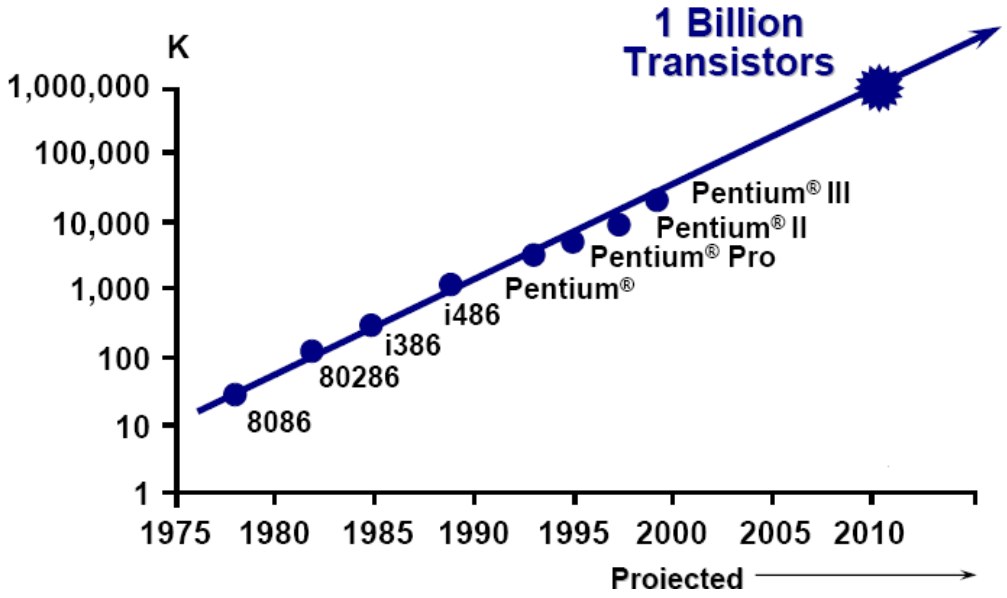


Figure 1.1: Number of transistors versus technology year [2].

Faster transistors were capable of switching faster thereby increasing the chip performance. However, increasing frequency of operation linearly increases the power dissipation in a chip.

1.2 Power Wall

Increase in power density within a chip due to higher frequency of operation is shown in Figure 1.2 for different technology nodes.

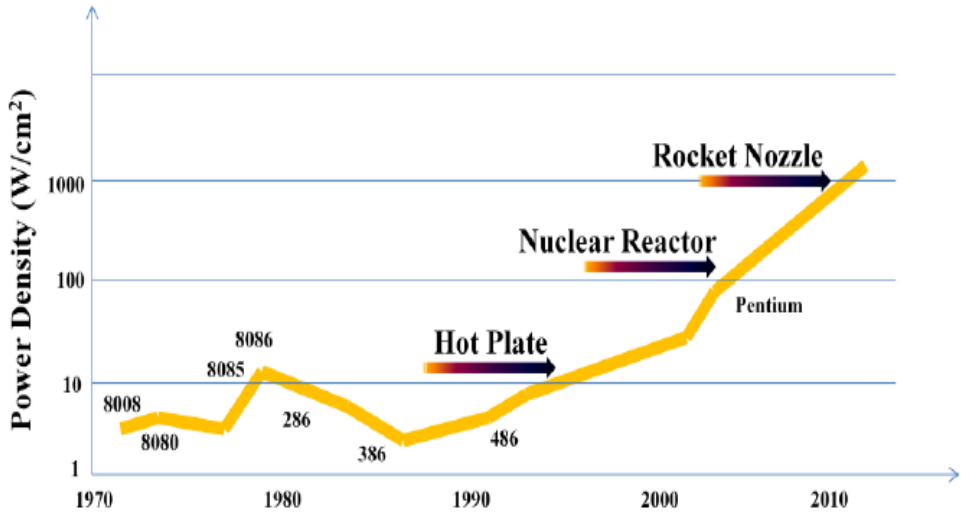


Figure 1.2: Power density versus technology year projection [3].

Extracting higher performance by increasing switching frequency is no longer an option due to absurdly high power densities within a chip. This problem is termed as the Power Wall, and power has become a first order design metric for chips. The semiconductor industry is hence in the midst of an important paradigm shift. To continue the aggressive performance growth, focus is turning towards many-core architectures.

1.3 Many-Core Era

In this technology many small cores are integrated on a single die. Performance is improved through parallelism. For the same power budget, total performance of the many-core system will be higher than that of the single processor [4]. Power dissipation in many-core chips can be maintained at acceptable levels through techniques such as fine grain power management [4]. Figure 1.3 shows the increase in number of cores per die as a function of technology year. It can be concluded that many-core technology is fast becoming popular and therefore it becomes important to evaluate the characteristics of its communication fabric.

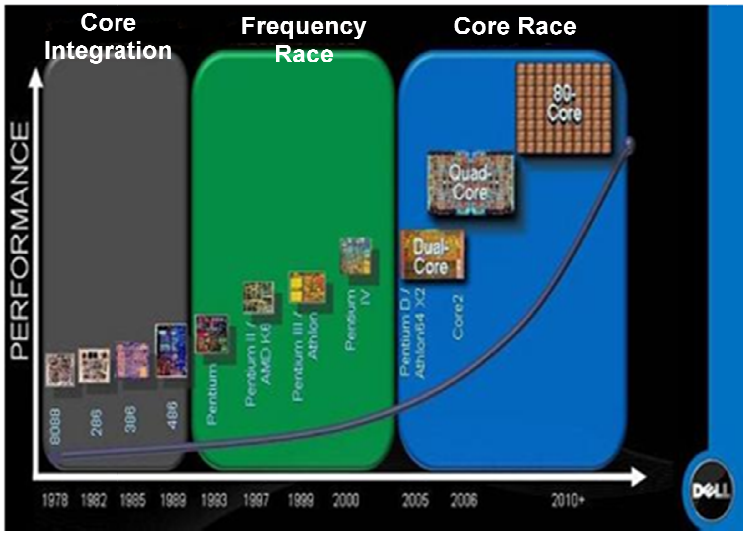


Figure 1.3: Performance versus technology node [5].

1.4 Network-on-Chip

The backbone of the many-core system is the network-on-chip (NoC) that connects all the cores together [4]. The major challenge for network-on-chip based global

interconnects in many-core architectures is to provide very large bandwidth (of the order of terabits/s) at low power and low latency [6], [7].

It can be noted that the many-core communication fabric is the network-on-chip, which comprises of copper wires and routers at each core. The topology of interconnection of various cores on a die could be different. Hence, to analyze the global interconnects for many-core architectures, it is imperative to look at the problem from an architectural as well as a technological perspective. In the processors of previous generations, which were single core, analysis of wires was decoupled from the architectural details.

1.5 Organization of Thesis

In Chapter 2, the global interconnects are optimized for maximum bandwidth density, minimum delay and minimum energy-per-bit. Unity aspect ratio wires are first considered and later, wires are optimized by imposing practical limitations. The optimal dimensions developed in Chapter 2 are then used in Chapter 3 to determine wiring area utilization, maximum core-to-core channel width and aggregate chip bandwidth for different network-on-chip topologies. Chapter 4 deals with analysis of 3D network-on-chip architectures. The conclusions and future work are presented in Chapter 5.

Chapter 2

Optimal Global Interconnects

2.1 Introduction

While integrated circuits are moving towards many-core architectures, no circuit-aware interconnect technology optimization methodology has been reported for such chips. To utilize a many-core chip to its full potential, low-latency ultra-high bandwidth inter-core interconnects are needed.

Previously, the width of gigascale global interconnects was optimized to achieve large bandwidth density and small latency simultaneously [8]-[12]. But these optimizations were not performed for network-on-chip (NoC) based systems; hence, the limitations imposed by routers were ignored. Also the impact of size effects on the resistivity of scaled copper wires was ignored in [8], [9].

This chapter will focus on the optimal design of inter-core interconnects in many-core architectures. The results have important implications for interconnect technology development for many-core chips. To highlight the importance of the limitations imposed by routers, interconnect dimensions with unity aspect ratio are first optimized based on the intrinsic properties of interconnects (Section 2.2) and later by taking into account the limitations imposed by the routers in a NoC (Section 2.3). In Section 2.4, global interconnects are then optimized by imposing practical limitations. The conclusions are summarized in Section 2.5.

2.2 Optimization Based on Intrinsic Properties of Interconnects

In this section, wire width is optimized without taking into account the limitations imposed by NoC and routers. The grain boundary and surface scatterings of a copper wire are modeled by the Maya-Shatzkes [13] and Fuchs-Sondheimer models [14], respectively. The reflection coefficient R and specularity parameter p for the models are assumed to be 0.5 [15]. Delay of a copper wire, τ, with optimal number of repeaters, in RC and RLC regions is modeled according to [16].

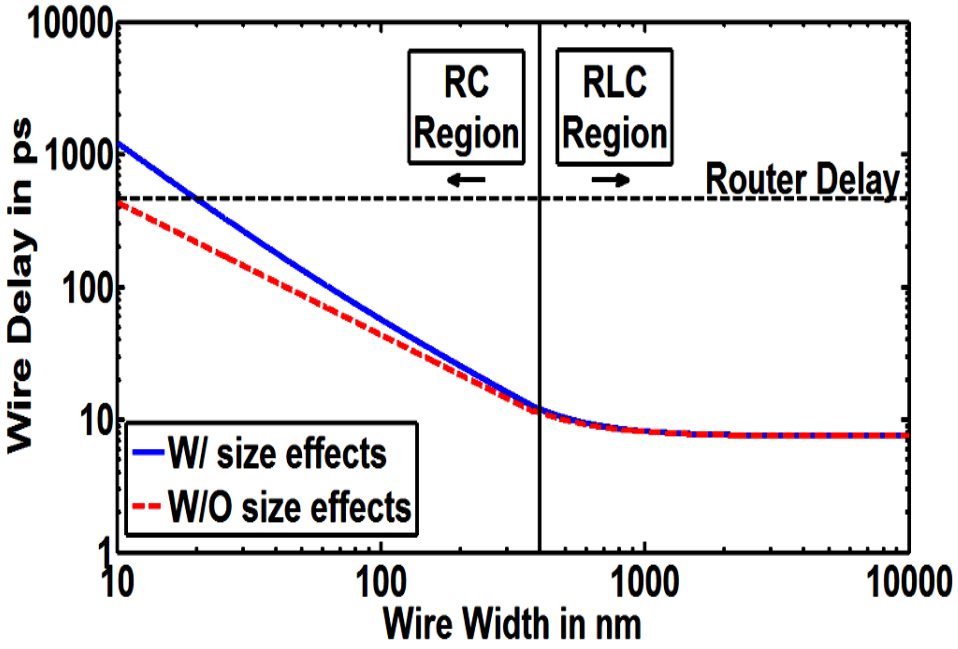


Figure 2.1: Copper wire delay, τ, with optimal number of repeaters, versus wire width, with and without copper size effects. Router delay for low network traffic is also shown for reference.

All cross-sectional interconnect dimensions are assumed to be equal and are changed proportionally as wire width increases. In this scenario, capacitance remains constant since all dimensions are changed proportionally and resistance decreases on increasing the width by which RC product decreases. This allows much more flexibility

in optimizing interconnect delay and bandwidth compared to the case that only wire width is changing. If the RC product becomes comparable or even smaller than the time of flight, interconnect operates in the RLC regime.

All technology parameters are projections of the International Technology Roadmap for Semiconductors (ITRS) [17] for the year 2015 (25nm node). Interconnect delay versus wire width is plotted for a wire length of 1.5mm in Figure 2.1 considering and ignoring size effects. For small dimensions, increasing wire dimensions lowers interconnect resistance and hence lowers the delay (RC region). For large dimensions, delay becomes time-of-flight limited and there is a diminishing return in delay (RLC region).

Bandwidth density (or data flux density, Φ_D) which is defined as the product of bandwidth and reciprocal interconnect pitch, represents the number of bits per unit time that can be transferred across a unit length bisectional line [8], [10]-[12], [18]. Bandwidth density variation versus interconnect width, with and without size effects is shown in Figure 2.2 (left vertical axis). It can be observed that bandwidth density is maximized at the boundary of RC and RLC regions where the size effects and inductive effects are both minimized. Previous optimizations did not consider size effects because of which the bandwidth density remained a constant in the RC region [8], [9]. Since bandwidth density and delay are equally important a more useful optimization metric therefore is Φ_D/τ [8], [10]-[12], [18]. This metric is plotted against wire width for an interconnect length of 1.5mm in Figure 2.2 (right vertical axis). This optimal width which maximizes Φ_D/τ is length independent and is slightly larger than the width at which bandwidth density is maximized.

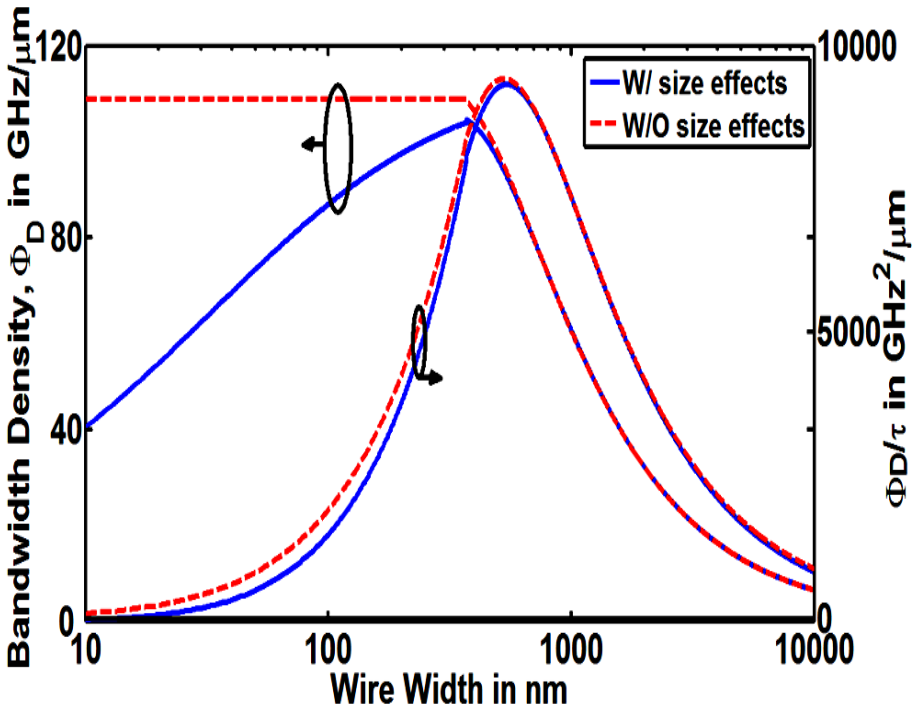


Figure 2.2: Bandwidth density and bandwidth density reciprocal latency product versus wire width for wire length of 1.5mm.

2.3 Unity Aspect Ratio NoC Interconnect Optimization

In this section, a first order optimization of inter-core interconnects is presented where interconnects with unity aspect ratio (W=S=T=H) are optimized for maximum bandwidth density and minimum delay. Since interconnect capacitance remains constant, energy-per-bit is not affected by this optimization.

2.3.1 Network-on-Chip (NoC)

Network-on-chips have been widely proposed as the interconnect fabric for many core architectures due to their performance, scalability and modularity [19]. Ring, 2D torus/mesh networks and their many variants are the candidate topologies [20]. In the rest of the chapter, a 2D mesh as shown in Figure 2.3 is used as an example topology. However, the analysis presented in this chapter can be applied to any network topology

provided accurate hop lengths (distance between two adjacent cores) and router latency details are available. In the example topology, each core has an on-chip 5-port router. The cores are connected through copper interconnects. For the year 2015, the ITRS projects a chip size of 391mm² [17]. A hop length of 1.5mm is therefore considered assuming an array of 12×12 cores on a chip.

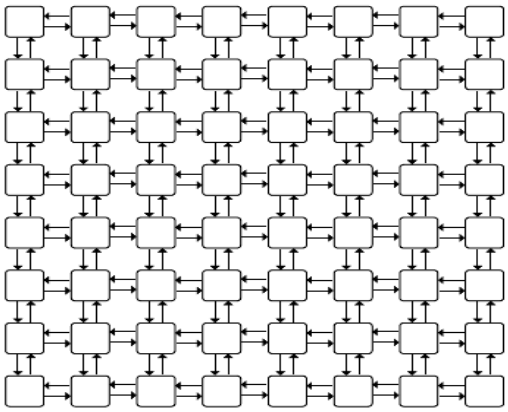


Figure 2.3: Two dimensional network-on-chip mesh topology [21].

NoC router forms a pivotal component of the NoC system and hence its delay needs to be taken into account. In the 2D mesh interconnected system a router is present in each of the cores. Delay of the router depends on the network data traffic and two extreme operating conditions are: no-load and full load with delays of 5 and 20 clock cycles, respectively [22]-[24]. As a best case scenario, no-load router delays are considered for optimization. However, optimization results remain the same for full-load operating condition.

2.3.2 Delay in Network-on-Chip (NoC)

In a NoC, total single-hop (transfer of data from one core to an adjacent core) delay, is the sum of wire delay and the delays of two routers. Two routers are considered as the data must be injected into the source router and then it is routed through the wire to the

destination node router.

Hence the total delay for *N* hops can be represented as:

$$\tau_{tot} = N\tau_{wire} + (N+1)\tau_{router} \tag{2.1}$$

The NoC router can latch data only at the rising or falling edges of the clock. Hence, the following limitation is imposed upon wire delay:

$$\tau_{wire} = \left[\frac{\tau}{\tau_{clk}} + 1\right]\tau_{clk} \tag{2.2}$$

where, [.] represents the integer part. For the technology year 2015, the projected clock frequency is 8.522 GHz [17] that corresponds to a clock-cycle of 117ps. τ is the copper wire delay as modeled in Section 2.2.

Bandwidth is the reciprocal of wire latency, τ_{wire} and is a constant for any number of hops because data can get pipelined by routers [23]. However, the overall delay for the traversal of the data from the source to destination increases linearly with the increase in number of hops. Corresponding bandwidth density is plotted for different hop lengths in Figure 2.4 (right vertical axis). The width at which bandwidth density is maximized is in the RC region in contrast to the optimal width obtained in Section 2.2.

2.3.3 Optimization Results and Discussion

Optimization metric Φ_D/τ_{tot}, for different hop lengths is also plotted in Figure 2.4 (left vertical axis). It can be clearly observed that that the optimal width is less than 100nm and is the same as the width at which bandwidth density is maximized. This is due to the clock delay limitation imposed by NoC and the delay of routers which is an order of magnitude larger than the wire delay. Optimal wire width approaches minimum dimension (25nm) as

hop length reduces (or number of cores on a die increases) and will therefore be minimum dimension limited for more than 1000 cores on a die for the technology year 2015 as shown in the inset plot in Figure 2.4.

The zigzag shape of the curves in Figure 2.4 at small wire widths is because routers can latch interconnect outputs only at the rising or falling edges of the clock signal. Thereby, the effective delays of interconnects are only integer multiples of clock cycles as described by (2.2). The local peaks in Figure 2.4 correspond to the wire widths at which copper wire delay, τ, is exactly a multiple of clock cycle. Slight increases in wire width beyond these points do not change the effective interconnect delay or bandwidth, and hence lowers bandwidth density.

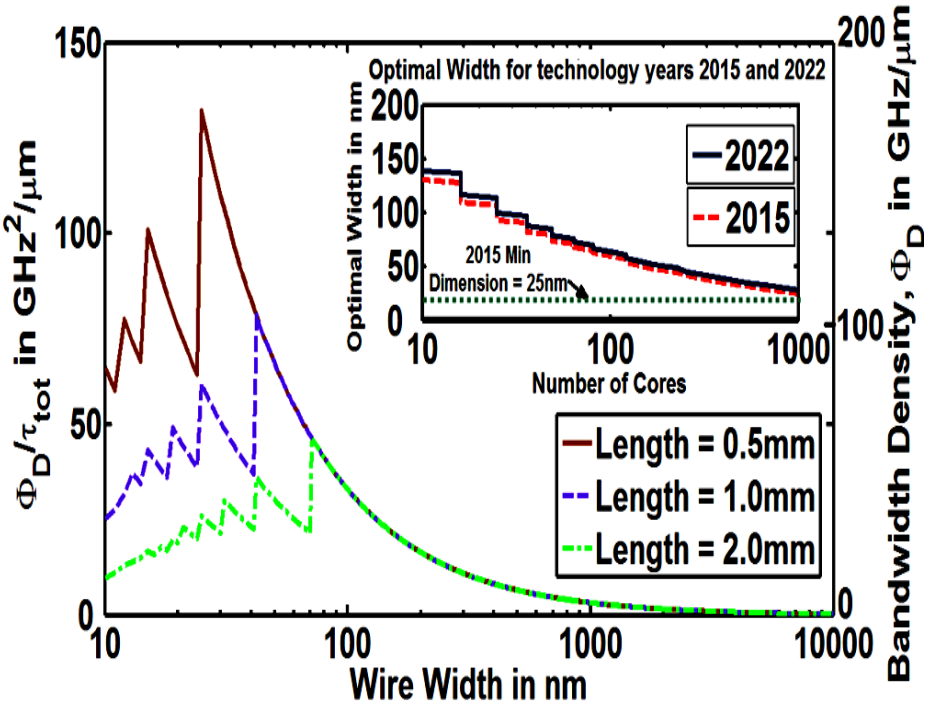


Figure 2.4: Bandwidth density and Φ_D/τ_{tot} versus wire width for different hop lengths. Inset plot shows the optimal wire width versus number of cores for technology year 2015 (25nm node). The optimal width approaches minimum dimension as the number of cores on a die increases.

The optimal width is in the deep RC region where inductance can be ignored. These results are in sharp contrast to that of previous optimizations where it was found that the global interconnect optimal wire width was many times larger than the minimum width and in the shallow RLC region. For a length of 1.5mm, the optimal wire width for the NoC based many-core system is 10 times smaller than that of a single core chip. Also, one can take a look at the significant drop in Φ_D/τ_{tot} caused by NoC limitations by comparing Figures 2.2 and 2. 4. For the same chip, on using a super- optimal width of $2W_{opt}$ worsens the bandwidth density by 50% with a small improvement in delay of 5%.

An expression for the optimal width, W_{opt}, is derived as:

$$W_{opt} = \left[3.5355\, lf\sqrt{\xi\beta\rho_b\varepsilon_o\varepsilon_r R_o C_o}\right]^{2/3} \tag{2.3}$$

where, l represents hop length, f is the clock frequency, ξ is a dimensionless quantity that depends on the geometry of the wire, β is a dimensionless coefficient that is a function of specularity parameter (p) and reflection coefficient (R) [25], parameter ρ_b is the bulk resistivity of copper, $\varepsilon_o\varepsilon_r$ is the dielectric permittivity and R_o and C_o are the output resistance and input capacitance of a minimum-size repeater, respectively. For unity aspect ratio, ξ= 6.05 [8].

A similar analysis has been conducted for all technology years till 2022 using the data provided in the ITRS. It is observed that the optimal width W_{opt} is a weak function of the technology year and is determined mainly by the number of cores or hop length.

2.4 Practical Optimization

In reality, aspect ratio of an interconnect within a chip is greater than one. As a result, the capacitance of interconnects would no longer be constant on varying its width. Therefore, in this section, width, pitch and number of repeaters for inter-core interconnects are

optimized to minimize energy-per-bit (E_b) and delay, and maximize bandwidth density, simultaneously. The interconnect height (H) and inter-layer dielectric thickness (T) are maintained constant. Interconnect self and mutual capacitances are calculated using the empirical models in [26]. The RC wire delay equation is given by [25]:

$$\tau_{wire} = \left(1.4 + 0.53\zeta + \frac{0.53}{\zeta}\right)\sqrt{R_o C_o rc}\,.\,l \tag{2.4}$$

Here ζ is the sub-optimal repeater insertion factor, given as the ratio of number of repeaters inserted to optimal number of repeaters and r and c respectively are per unit length interconnect resistance and capacitance. The total repeater capacitance and the number of repeaters are given respectively as [8]:

$$C_{rep} = \zeta 0.75\ C_{int} \tag{2.5}$$

And

$$\text{Number of Repeaters, k} = \zeta\ \sqrt{\frac{0.4\ R_{int}C_{int}}{0.7\ R_o\ C_o}} \tag{2.6}$$

All technology parameters are projections of the ITRS [20] for the year 2015 (25nm node). Bandwidth density is plotted against pitch for different hop lengths in Figure 2.5. It is worthwhile to note that the minimum pitch for global interconnects in 2015 is 75nm [17]. For each point in Figure 2.5 width and spacing of interconnects are optimized to maximize bandwidth density and minimize delay and energy-per-bit.

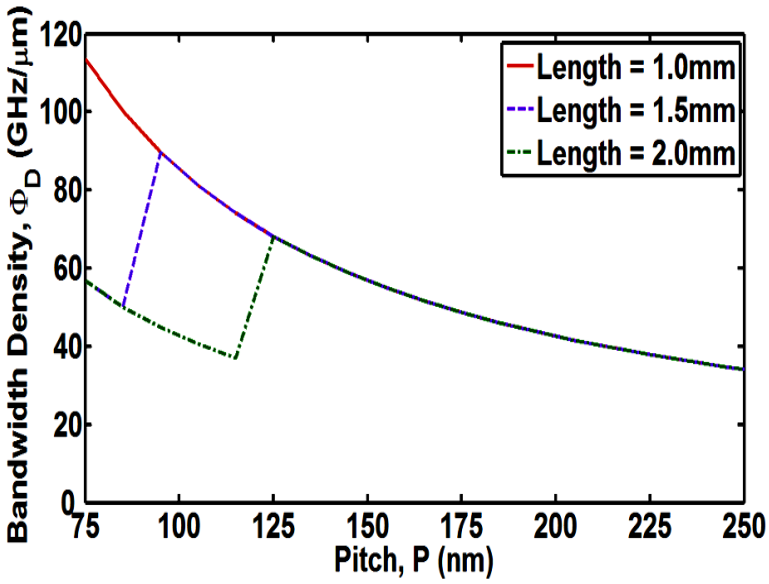


Figure 2.5: Bandwidth density versus pitch for different hop lengths in technology year 2015.

Since the capacitance would no longer remain constant for different width, the energy-per-bit variations are as shown in Figure 2.6.

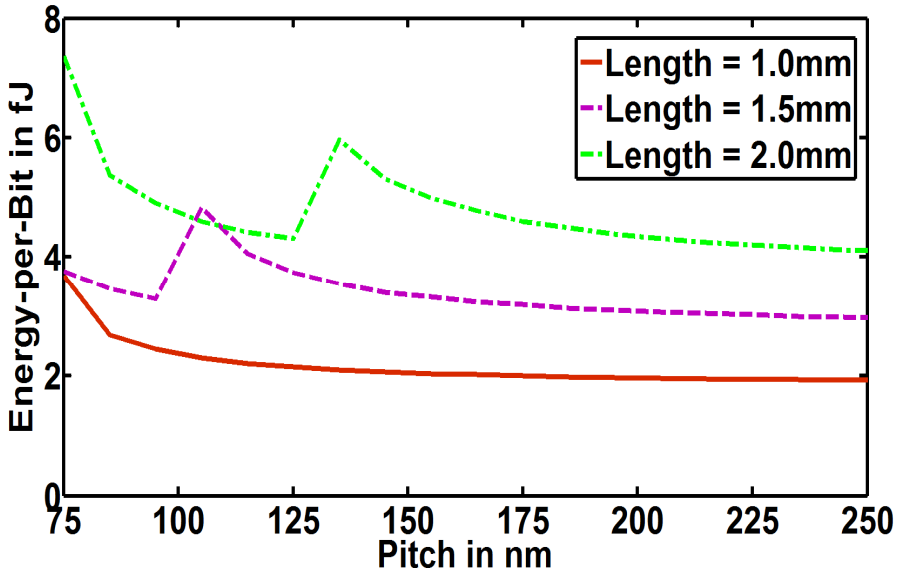


Figure 2.6: Energy-per-bit versus pitch for different hop lengths in technology year 2015.

Since bandwidth density, delay and energy are important, a new metric $\Phi_D/(E_b\tau_{tot})$ is defined to determine the optimal dimensions. This metric is plotted versus pitch for different hop lengths in Figure 2.7. Again in Figure 2.7 for every pitch, the optimal width has been calculated.

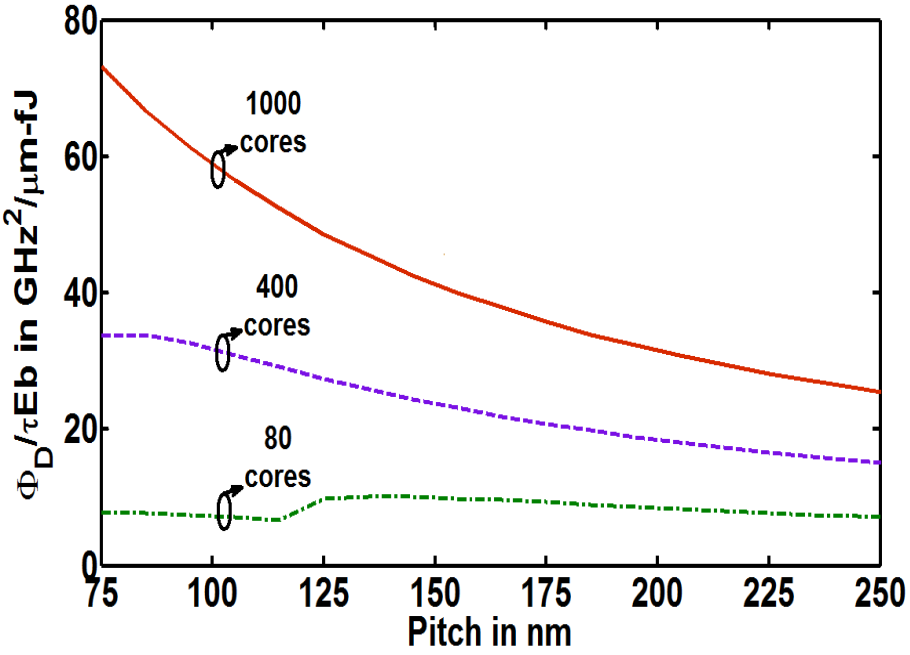


Figure 2.7: Product of bandwidth density, reciprocal latency and energy-per-bit versus pitch in technology year 2015 for various hop lengths. 1000, 400 and 80 cores on a die correspond to hop lengths of 0.6mm, 1mm and 2mm, respectively.

Optimal pitch is where the metric $\Phi_D/(E_b\tau_{tot})$ is maximized. From Figure 2.8 it can be noted that the optimal pitch becomes minimum pitch limited for more than 500 cores. Once again to see the effectiveness of this optimization one can consider a 1000 core chip. On using double the optimal pitch (150nm) energy-per-bit improves by 11%, delay improves by just 0.5%, but bandwidth density worsens by 52%.

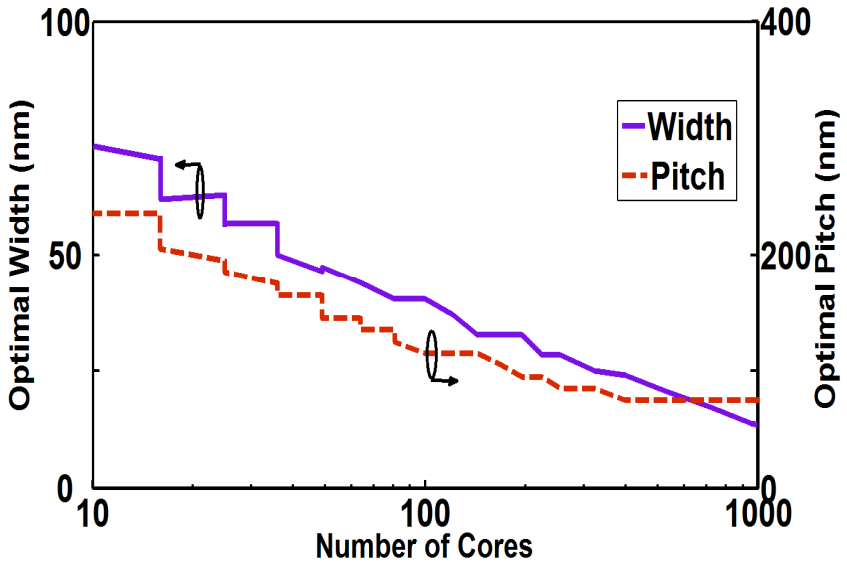


Figure 2.8: Optimal width and pitch for different cores on a die in the technology year of 2015. For every value of pitch, the width and spacing are optimized for maximizing $\Phi_D/(E_b\tau_{tot})$.

In previous single core interconnect optimizations, repeater insertion factor was taken to be 50% of the optimal number [27]. This is attributed to the minimal improvement in wire delay on increasing ζ above 0.5. It is demonstrated in Figure 2.9.

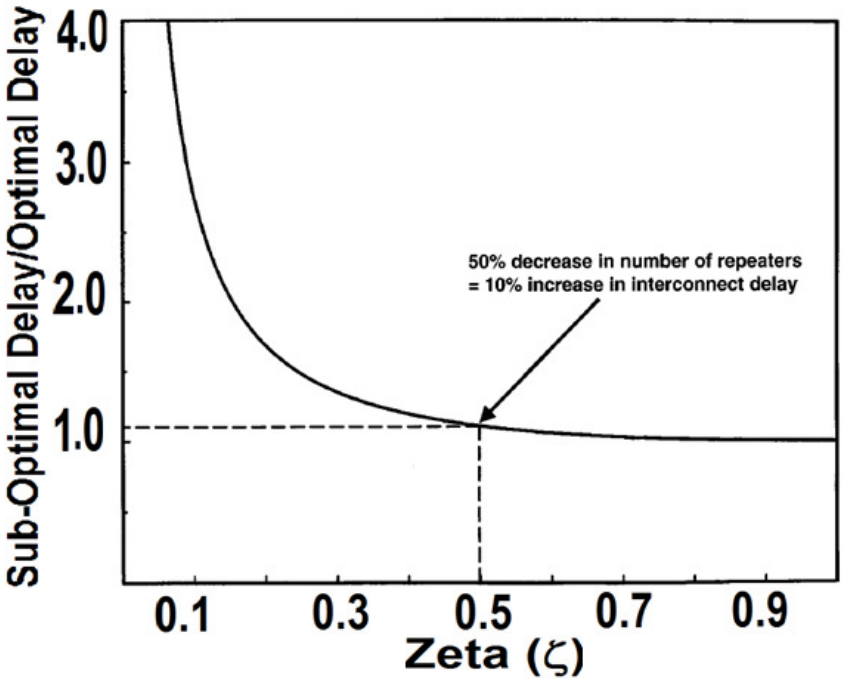


Figure 2.9: Sub-optimal delay over optimal delay versus repeater insertion factor, ζ [27].

Therefore, until this point, the repeater insertion factor, ζ was assumed to be 0.5 which gives the opportunity for further optimization. Figure 2.10 shows the metric $\Phi_D/(E_b\tau_{tot})$ versus number of repeaters. From Figure 2.10, it can be clearly noted that using a ζ less than 0.5 results in further improvement of $\Phi_D/(E_b\tau_{tot})$. It is observed that, using 8 repeaters instead of 22, results in 16% improvement in energy-per-bit with just 2% delay penalty and no impact on bandwidth density. It is worthwhile to note that for a 1000 core chip, the optimal dimensions are minimum pitch and minimum width limited.

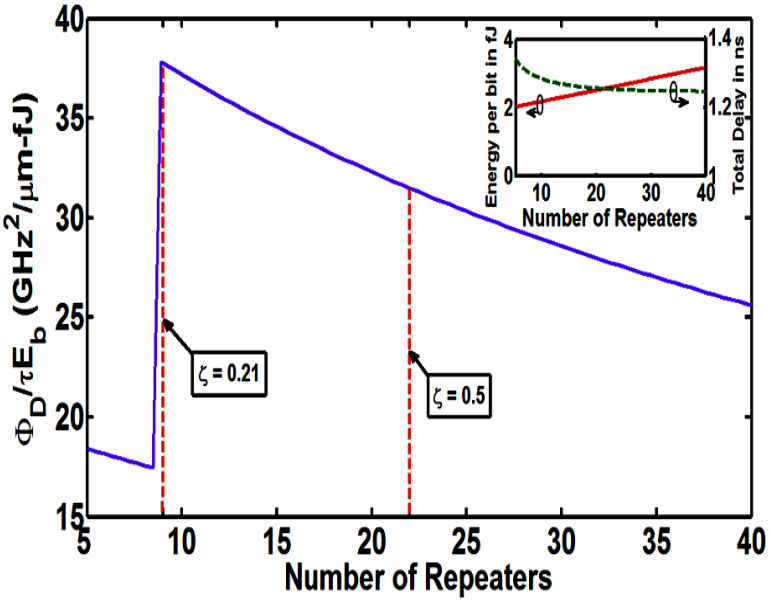


Figure 2.10: $\Phi_D/(E_b\tau_{tot})$ versus number of repeaters in technology year 2015 for a 1000 core chip in 2015. Inset plot shows the variation of energy-per-bit and total delay versus number of repeaters. For every ζ, the pitch, width and spacing are optimized for maximizing $\Phi_D/(E_b\tau_{tot})$.

Extending the above repeater insertion analysis to different number of cores, it is determined that repeaters can be inserted less frequently than required for single core chips with minimal impact on performance. This is due to delay of routers, which are orders of magnitude larger than wire delays. The required repeater insertion for different cores is depicted in Figure 2.11.

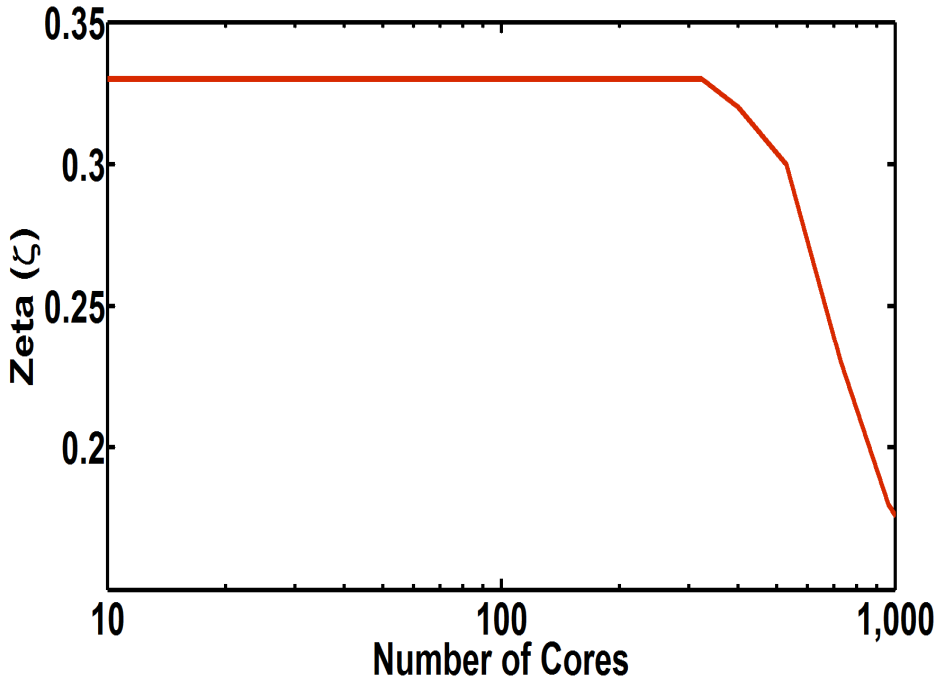


Figure 2.11: Optimization of repeater insertion factor, ζ for different number of cores per die. The die area is maintained constant at 391 mm² [17].

2.5 Conclusions

In this chapter, for the first time a circuit-aware interconnect technology optimization is performed for a network-on-chip in many-core architecture. First, global interconnects with unity aspect ratio are optimized to achieve maximum bandwidth density and minimum delay simultaneously. Optimal wire-width, W_{opt}, for a 144-core chip in the technology year 2015 is found to be more than 10 times smaller than previous optimization results where router latency, reduced interconnect length and size effects of copper wires were ignored. For the same chip, on using a super- optimal width of $2W_{opt}$ worsens the bandwidth density by 50% with a small improvement in delay of 5%. The optimal wire width does not vary significantly with technology year, but is a strong function of number of cores on a die and the frequency of operation. These results have

important implications for interconnect process development for future technology nodes and also for benchmarking emerging interconnect technologies.

Second, practical limitations, such as, aspect ratio greater than unity are imposed on global interconnects to optimize for maximum bandwidth density, minimum delay and energy-per-bit simultaneously. For a 1000 core chip in 2015, the optimal dimensions are minimum pitch and minimum width limited. The optimal number of repeaters is around 0.3 times the optimal number of repeaters found based on intrinsic delay of interconnects. This along with width and spacing optimization achieves a crucial 16% reduction in energy-per-bit.

Chapter 3

2D Interconnect Network Analysis

3.1 Introduction

The many-core communication fabric, network-on-chip, comprises of copper global wires and routers at each core. There are different topologies of network-on-chip and the architecture of router varies with each topology. Hence, to analyze the global interconnects for many-core architectures, it is imperative to look at the problem from an architectural as well as a technological perspective. In the processors of previous generations, which were single core, analysis of wires were decoupled from the architectural details. Recently, the global interconnects in many-core architectures were optimized to achieve high bandwidth density and minimum delay simultaneously [28].

In this chapter, channel width, which represents the number of wires connecting two cores within a many-core chip, is determined by using optimal dimensions of the global interconnects as calculated from [28]. The channel width in a many-core chip is then studied for different network-on-chip topologies for a given technology node. Router area within each core is determined as the main limiting factor to increasing channel widths.

Mesh topology is first considered for its simplicity. In Section 3.2, optimal wire dimensions are determined for global wires in mesh topology, to maximize bandwidth density and minimize delay, simultaneously. This is followed by, energy analysis for various channel widths. Wire as well as router power dissipation are ascertained and an optimal channel width is chosen to minimize energy-per-bit. The same analysis is then repeated for other topologies, such as, concentrated mesh, flattened butterfly and

concentrated flattened butterfly in Section 3.3 and the results are compared. The conclusions are summarized in Section 3.4.

3.2 Mesh Topology

In this section, global wire width for a mesh topology, as shown in Figure 3.1, is first optimized taking into account the limitations imposed by NoC and routers. Energy-per-bit in mesh topology is then determined by considering routers and wires power dissipation.

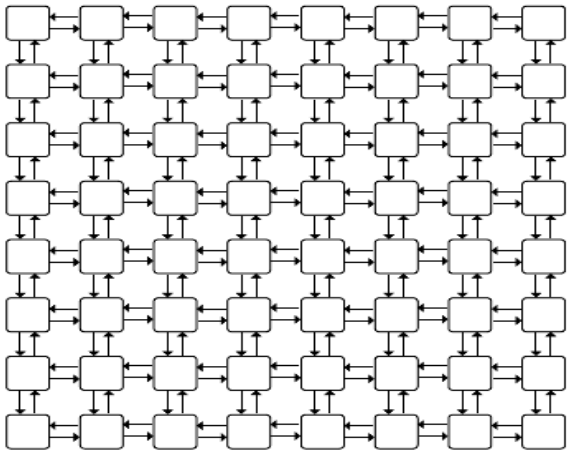


Figure 3.1: Mesh topology [21].

3.2.1 Optimal Wire Width

All cross-sectional interconnect dimensions are assumed to be equal and are changed proportionally as wire width increases. In this scenario, capacitance remains constant since all dimensions are changed proportionally and resistance decreases on increasing the width by which RC product decreases. This allows much more flexibility in optimizing interconnect delay and bandwidth compared to the case that only wire width

is changing. If the RC product becomes comparable to or smaller than the time of flight, interconnect operates in the RLC regime.

All technology parameters are projections of the International Technology Roadmap for Semiconductors (ITRS) [29] for the year 2012 (32nm node). However, the analysis can be repeated for other technology years as well. The 32nm technology node is considered to match with results of ORION 2.0 [30], a network-on-chip simulator which will be described later.

Optimal wire width at which bandwidth desnity is maximized and delay is minimized simultaneously, is plotted for various number of cores in Figure 3.2. It is observed that the optimal width W_{opt} is a weak function of the technology year and is determined mainly by the length of wires.

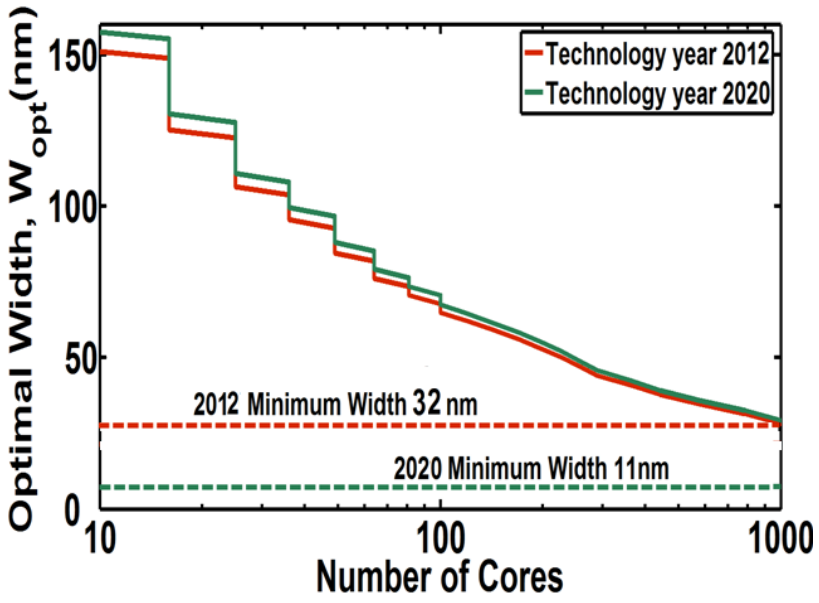


Figure 3.2: Optimal wire width versus number of cores for mesh topology. Die area is maintained constant at 413 mm² for ITRS 2012 [29].

3.2.2 Energy-per-Bit

It is important to study the energy-per-bit in network-on-chips as power has emerged as a first order design metric for modern multiprocessors. In network-on-chips, energy-per-bit has two main components: wire and router components.

Energy-per-bit due to wires depends on the number and size of repeaters, and wire capacitance. In the analysis, for simplicity, it is assumed that width, height, thickness and spacing of wires are equal and are changed proportionately. Also, optimal number of repeaters are inserted along each wire.

NoC routers in the mesh topology are considered to have five stages: input buffer, route computation, switch allocation, crossbar and output buffer. The router area and power dissipation vary with the channel width. ORION 2.0, a power simulator for interconnection networks, is used to determine the power dissipation and area of routers for different channel widths [30]. The ITRS 32nm technology node is set as the technology year for ORION 2.0. Router architecture and channel widths are given as input to ORION 2.0. Corresponding energy-per-bit and area of routers are then calculated by ORION 2.0.

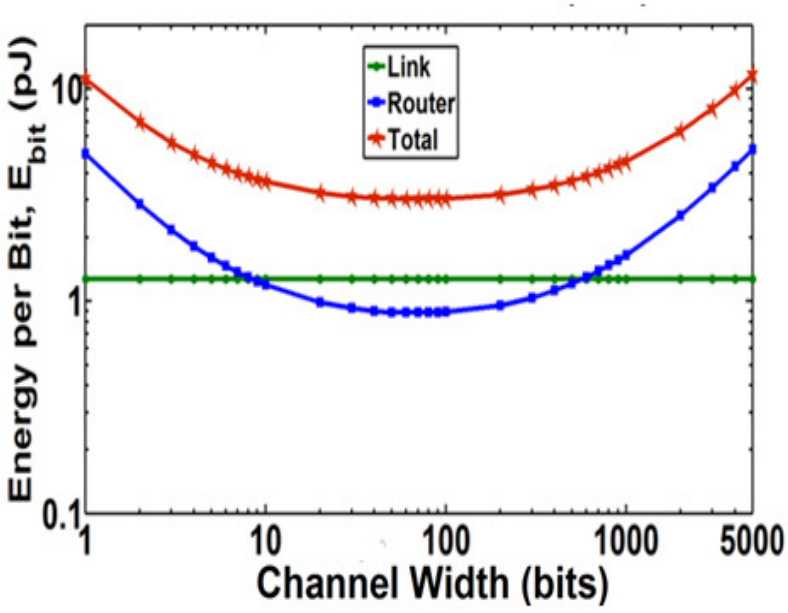


Figure 3.3: Energy-per-bit for one hop versus channel width for a 9-core (3×3) 2D mesh.

Figure 3.3 shows the variation of energy-per-bit for a 9-core mesh based NoC. Energy-per-bit is calculated for one hop, consisting of two routers and one wire power dissipation. The wire energy-per-bit remains constant when channel varies because the capacitance of the wires does not change.

One can observe from Figure 3.3 that the router energy-per-bit decreases at first, reaches a minima before increasing for larger channel widths. This non-linearity can be attributed to the characteristics of components within the router. Route computation and switch allocation stages are found to be dominant at smaller channel widths and crossbar and buffer stages at larger channel widths.

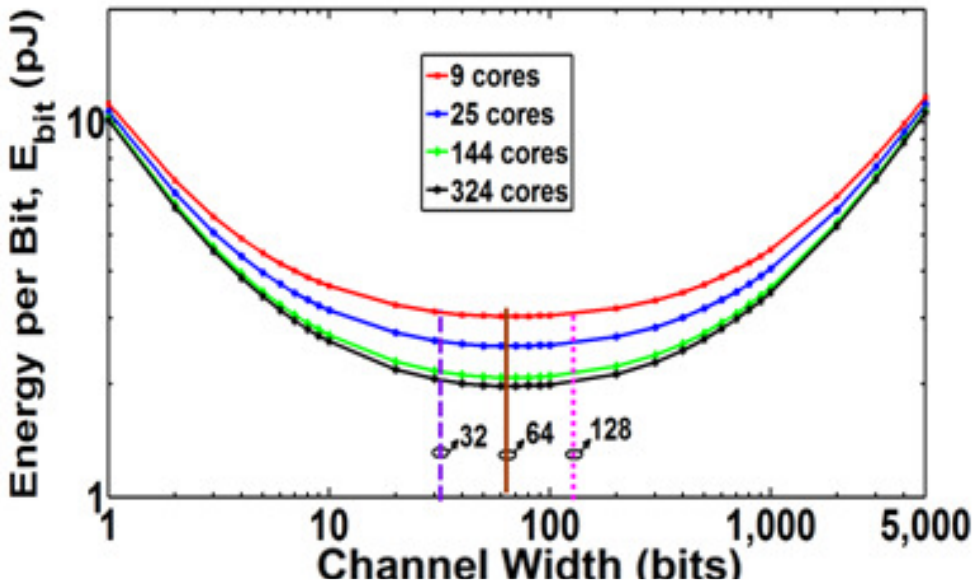


Figure 3.4: Energy-per-bit versus channel width for different number of cores per die. Die area is maintained constant at 413 mm² for ITRS 2012 [29].

The same analysis is repeated for higher number of cores keeping die area the same. The energy-per-bit variation is shown in Figure 3.4. The wire length between two cores reduces for higher number of cores for same die area. As a result the wire

capacitance decreases which reduces the interconnect component of the energy-per-bit. Since the router component of energy-per-bit remains the same and the wire component of energy-per-bit decreases as the number of cores increases, the channel width that minimizes energy-per-bit, remains independent of number of cores per die. Using 64 bits would result in the lowest energy-per-bit.

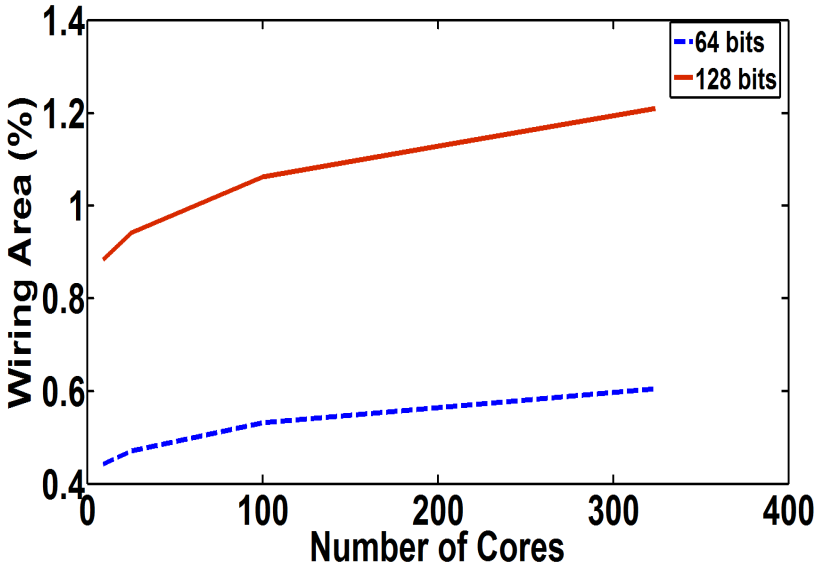


Figure 3.5: Wiring area expressed as a percentage of available wiring area (two orthogonal metal levels) for channel widths of 64 and 128 bits versus number of cores. Die area is maintained constant at 413 mm² for ITRS 2012 [29].

There is a small increase in energy-per-bit, of 0.7%, on increasing the channel width from 64 bits to 128 bits. Hence, it is worthwhile to compare wiring area utilization of the two channel widths. It is assumed that the inter-core wires are routed in two orthogonal metal levels of the chip. Wiring area expressed as a percentage of available wiring area (two orthogonal metal levels) for channel widths of 64 and 128 bits is shown in Figure 3.5. Optimal wire dimensions are first calculated for each number of cores. The wiring area is then calculated by adding up widths of 64 and 128 wires within an inter-core channel.

From Figure 3.5 it is evident that inter-core channel utilizes small wiring area. One may increase the channel width to increase aggregate chip bandwidth as depicted in Figure 3.6. However, this comes at the penalty of increased energy-per-bit as alluded to in Figure 3.4.

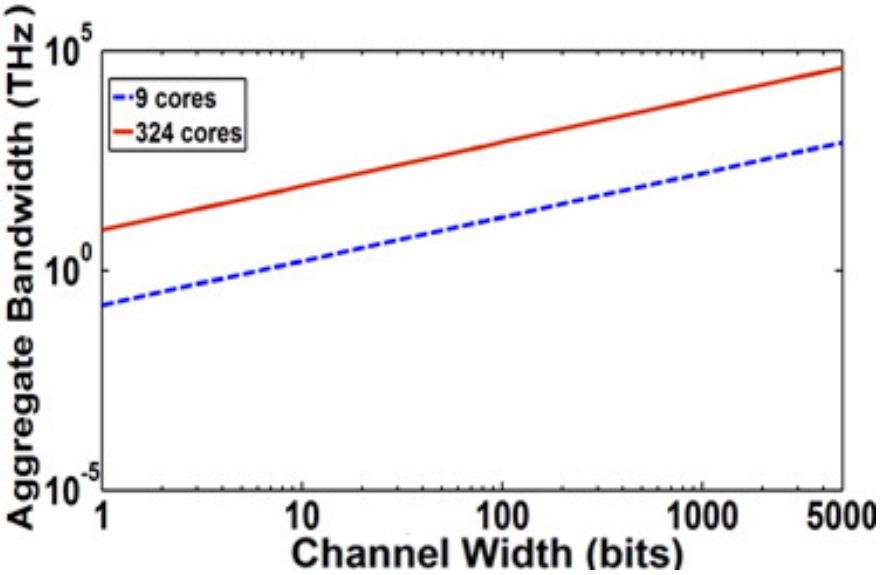


Figure 3.6: Aggregate bandwidth versus channel width for mesh based many-core chip.

Router area is directly dependent on the channel width. Figure 3.7 shows the increase in router area versus channel width. It also shows the area of one core in 9-core and 324-core chips. The number of cores, 9 and 324 are chosen to form simple 3×3 and 18×18 mesh based many-core chips. In a 9-core chip, if a channel width of 700 bits would be implemented, the router would occupy one whole core. This is impractical as routers should occupy a small fraction of cores.

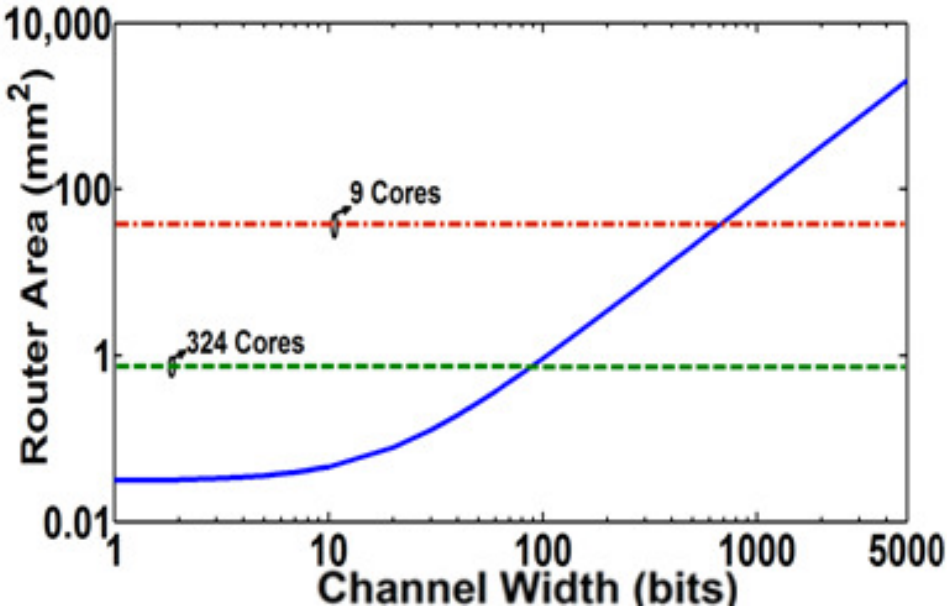
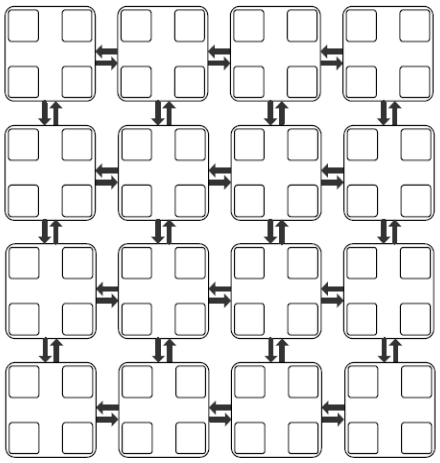


Figure 3.7: Router area versus channel width. ORION 2.0 was used to determine area of router. Area of a single core in a 9 core and 324 core chip is also shown for comparison. Die area is maintained constant at 413 mm².

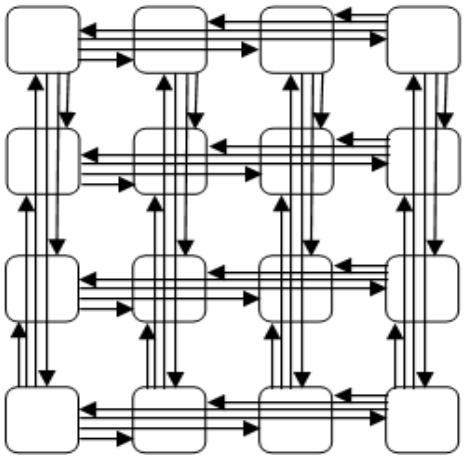
An important conclusion can be arrived at from the above discussion. Even though larger channel widths can be implemented due to the large available wiring area, area occupied by router within a core will limit the channel width. Hence, the aggregate bandwidth in many-core chips is not limited by wires, instead it is limited by router area. This is vastly different from single core chips where there were no routers and the aggregate bandwidth was determined by maximum number of wires within the chip.

3.3 Comparative Study of NoC Topologies

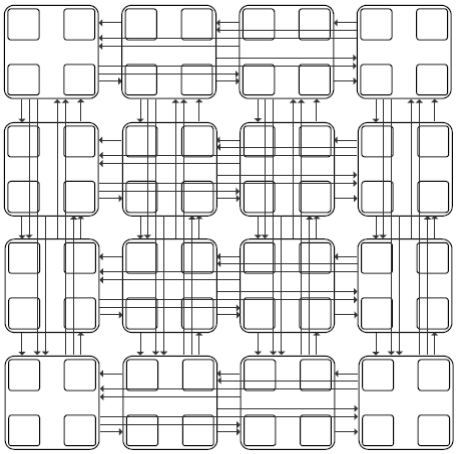
In this section, the analysis performed in Section 3.2 is extended to other network-on-chip topologies, such as, concentrated mesh, flattened butterfly and concentrated flattened butterfly. They are shown in Figure 3.8.



(a) Concentrated Mesh



(b) Flattened Butterfly



(c) Concentrated Flattened Butterfly

Figure 3.8: Network-on-chip topologies considered in this Section [21].

Concentration of a particular topology reduces the number of hops to transfer data from one end of the chip to another. The concentration factor for concentrated mesh and concentrated flattened butterfly is assumed to be a reasonable size of 4. However, the same analysis can be extended for other concentration factors.

It was determined in Section 3.2 that the optimal wire width was proportional to wire length. One can observe from Figure 3.8 that the wire lengths are different for mesh and flattened butterfly topologies. As the number of core increases, it would have an

opposite effect on the wire lengths of mesh and flattened butterfly topologies. The wire length in the mesh topology reduces as the core size would reduce keeping die area the same for large number of cores. However, in flattened butterfly topology, the length of longest wire would slightly increase as the number of cores increases.

Optimal wire width, the wire width at which wire delay is equal to one clock cycle, for mesh and flattened butterfly topologies is shown in Figure 3.9. It can be noted from Figure 3.9 that, optimal wire width for flattened butterfly topology increases with increase in number of cores and for mesh topology it decreases.

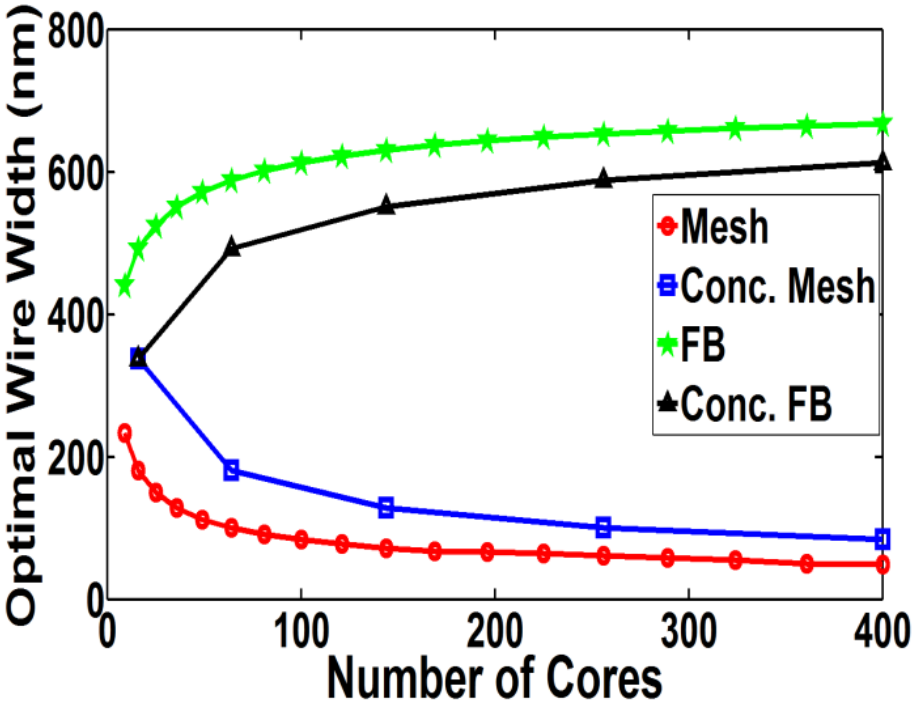
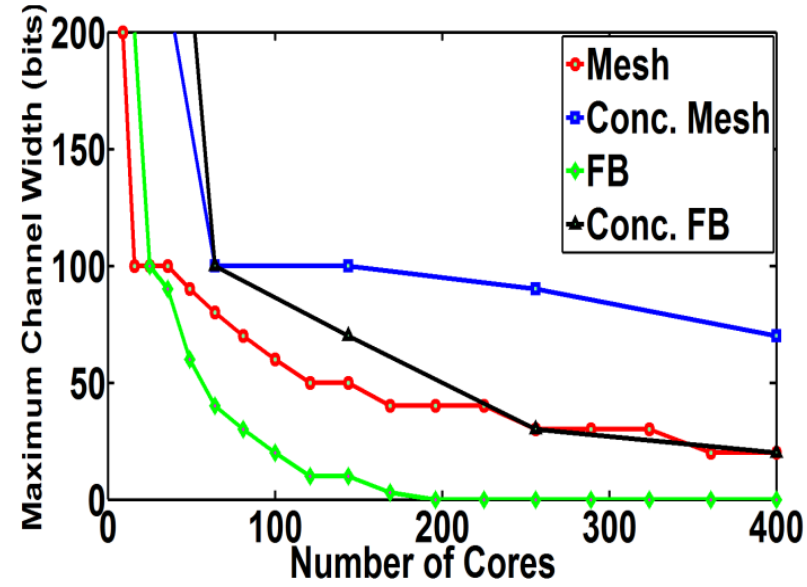


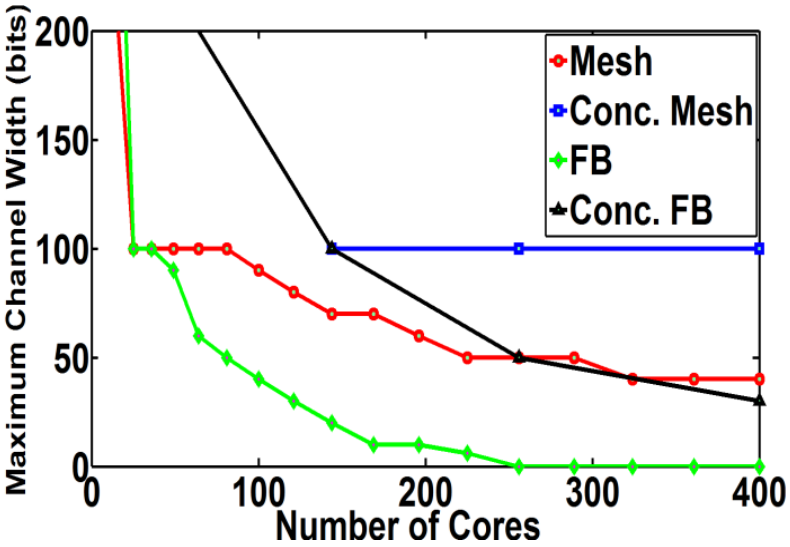
Figure 3.9: Optimal wire width versus number of cores for different topologies. Die area is maintained constant at 413 mm² for ITRS 2012 [29].

In Section 3.2, router area was determined to be a major bottleneck to increasing channel width. One way to analyze the problem would be to fix router area and then determine the maximum possible channel width for each topology. Figure 3.10 depicts

the maximum channel width for each topology by restricting router area as a fraction of single core area.



(a)



(b)

Figure 3.10: Maximum channel width versus number of cores on a fixed die area for different topologies. (a) Router area is assumed equal to 10% of single core area. (b) Router area is assumed equal to 20% of single core area.

It can be observed from Figure 3.10 that for flattened butterfly topology, channel width reduces to zero for 196 or more cores. If router area is increased to 20% of core area, then the maximum channel width could be stretched to 256 cores. The other topologies also reduce the maximum channel width as number of cores increase, however they are more scalable than flattened butterfly. Among the other topologies, concentrated mesh provides best scalability in terms of channel width.

With the knowledge of channel width and optimal wire dimensions, it is possible to evaluate the wiring demand for each topology. Figure 3.11 shows the variation of wiring area utilized against number of cores for various topologies. A large difference in wiring area can be observed between different topologies. This is due to the opposite effect on optimal wire dimensions on increasing the number of cores.

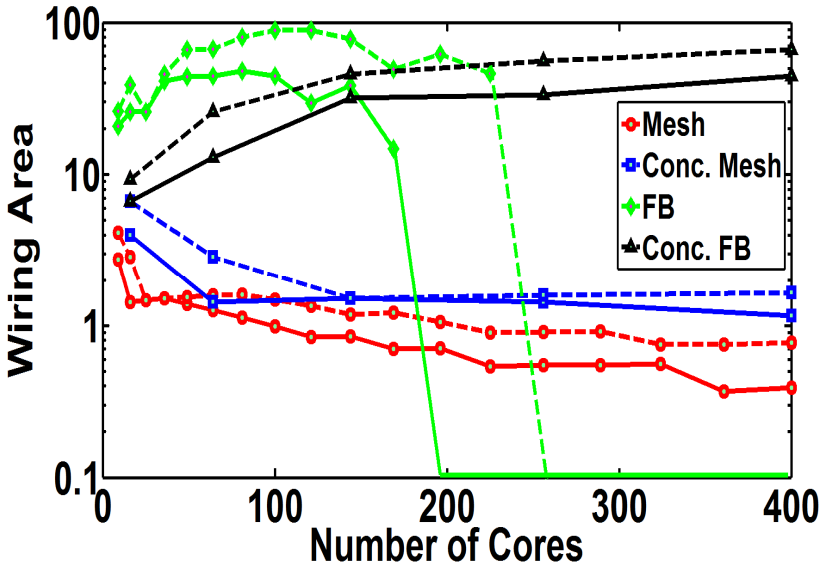


Figure 3.11: Wiring area versus number of cores on a fixed die area for different topologies. Dashed line indicates router area is 20% of single core area. Solid line indicates router area to be 10% of single core area. The top two orthogonal pair of metal levels are assumed available for routing.

One can observe from Figure 3.11 that flattened butterfly provides the largest wiring area utilization (more than 70%). Mesh topology highly under utilizes the

available wiring area (less than 5%). However, for flattened butterfly topology, the channel width reduces to zero as the number of cores increases. Increasing the concentration helps to scale well for higher number of cores.

Flattened butterfly topologies have the advantage of bypassing routers thereby reducing the number of hops required to transport data across the die. In mesh topology, the data would need to go through a series of hops before reaching the destination. However, since the number of ports for mesh is the same (five ports) for any number of cores, the channel width can be higher than flattened butterfly. In flattened butterfly topology, the number of ports per router would increase with the number of cores, as each core is connected to every other core in its dimension. Since both bandwidth and hops are important, a new metric aggregate bandwidth-hops is used to compare these topologies. For each channel, bandwidth-hop is the product of channel bandwidth and its number of hops.

Figure 3.12 shows bandwidth-hops metric versus a number of cores for various topologies. The advantage of flattened butterfly is over shadowed by its long and hence wide interconnects. Wider interconnects along with larger router areas contribute to smaller channel widths which reduces the channel bandwidth in flattened butterfly topology.

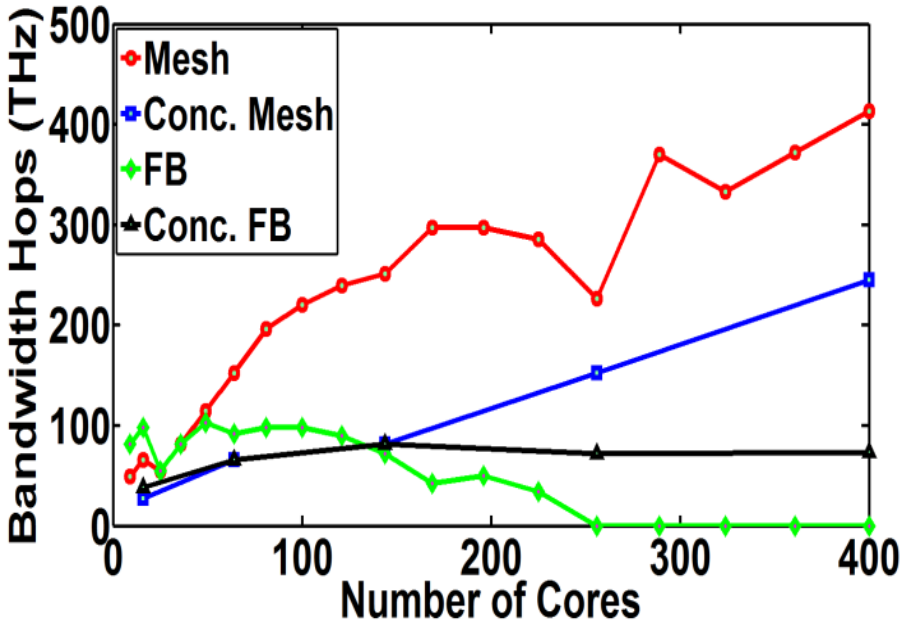


Figure 3.12: Bandwidth hops versus number of cores on a fixed die area for different topologies. Router area is assumed equal to 20% of single core area.

It is worthwhile to compare the maximum bisection bandwidth for each of the topologies. Flattened butterfly would have the highest bisection bandwidth as many channels cluster near the middle of the chip. It can be observed from Figure 3.8 (b) and (c). This advantage of flattened butterfly can be seen in Figure 3.13 which plots maximum bisection bandwidth versus number of cores. However, the advantage of flattened butterfly soon falls off due to the increase in number of ports per router with increase in number of cores.

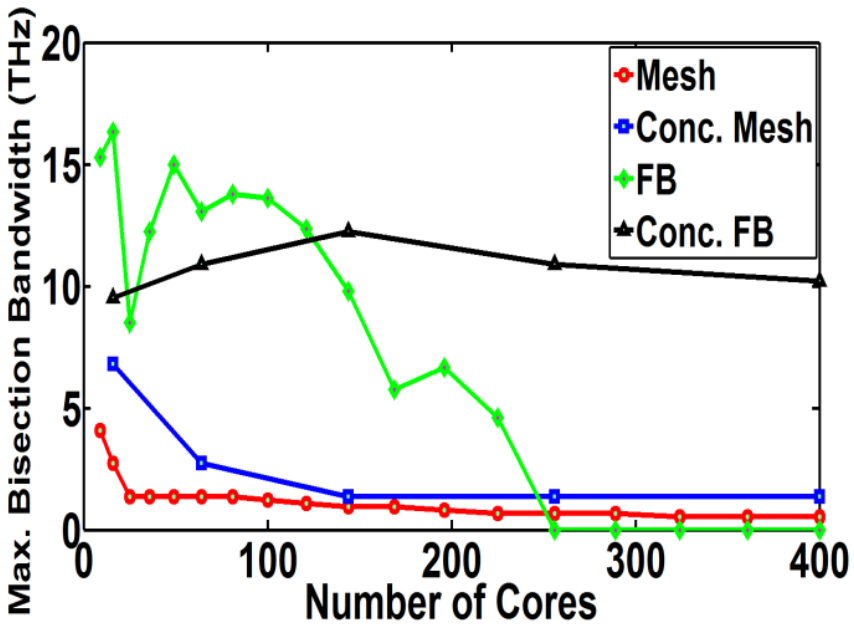


Figure 3.13: Maximum bisection bandwidth versus number of cores on a fixed die area for different topologies. Router area is assumed equal to 20% of single core area.

As mentioned earlier, the main advantage of a flattened butterfly topology is the possibility of bypassing routers when transporting data from one end to another. In Figure 3.14 the worse case delay is plotted for mesh, concentrated mesh and flattened butterfly topologies. Worse case delay is the time taken to transport data from one corner of the die to diagonally opposite corner. For mesh and concentrated mesh topologies this will involve passing through many hops and routers. As the number of cores on a die increases, the number of routers in the worse case path increases which would then increase the delay. It is assumed here that each hop has a delay equivalent to one clock cycle and each router contributes 5 clock cycles of delay. In the flattened butterfly topology the worse case delay comprises of two hops through three routers irrespective of number of cores. Hence the worse case delay remains the same on increasing the number of cores within a die.

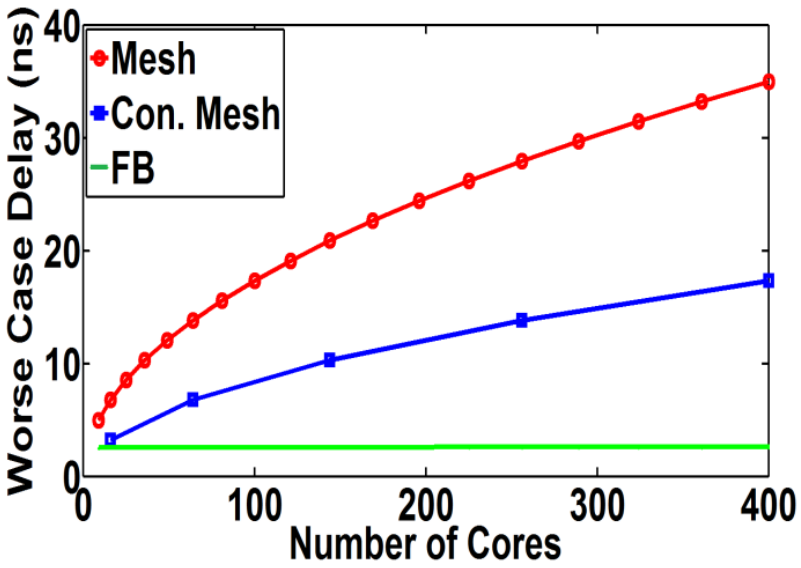


Figure 3.14. Worse case delay versus number of cores on a fixed die area for different topologies. No router area restriction is applied, i.e., a channel width of at least one bit is possible for flattened butterfly topology. If a restriction in router area (of 20% of core area) is imposed then the green line would not be valid beyond 256 cores as channel width would drop to zero.

3.4 Conclusions

In this chapter, for the first time a technology-cum-architecture aware network-on-chip interconnect optimization is performed for many-core chips. Optimal wire dimensions are determined for mesh, concentrated mesh, flattened butterfly and concentrated flattened butterfly topologies. Then using ORION 2.0, router area and router energy-per-bit are deduced. It is shown that router area is the main bottleneck to achieve larger core-to-core bandwidth. Flattened butterfly and concentrated flattened butterfly topologies have the best wiring area utilization, maximum bisectional bandwidth and the lowest worse case delay. However, these advantages diminish for larger number of cores due to the rapid increase in router area. Mesh and concentrated mesh topologies are found to be more scalable in terms of number of cores than flattened butterfly and concentrated flattened butterfly topologies.

Chapter 4

3D Interconnect Network Analysis

4.1 Introduction

Three-dimensional (3D) integrated chip designs are now widely acknowledged as the future of chip design. Some of the key technologies needed to enable 3D chip stacking include wafer bonding, wafer thinning, chip alignment, and fabrication of Through Silicon Vias (TSVs) with high-density lead-free solder interconnects [31].

Three-dimensional integration might alleviate a number of immediate problems faced by system architects. The first problem is the infamous "memory wall." Historically, processor performance has improved by about 60% per year, whereas the corresponding improvement in memory access time has been less than 10% per year. This gap is a major factor that limits overall system performance improvement and is just one of several factors comprising the memory wall. This effect can be even more pronounced with the increasing number of cores in the many-core context [32]. Three-dimensional technology can enable the integration of memory layers onto the processor chip and can thereby eliminate the slower and higher-power off-chip buses to that memory by replacing them with high-bandwidth and low-latency vertical interconnections to the memory layers.

The 3D integration provides several other major advantages: it addresses the critical interconnect bottleneck in today's chip designs. In 3D, the silicon footprint in each layer is much smaller than a 2D implementation of the system. This is intuitive, as a 3D design can be thought of as a folding of a 2D design in to multiple layers. The vertical interconnects connecting two silicon layers are short and typically have lower delay and

power dissipation than a corresponding 2D interconnect. Thus, the wire delay and wire power dissipation will be lower in a 3D design. The other advantages of 3D integration include smaller form factor and ease of integration of diverse technologies in a single chip [33].

Networks on Chips (NoCs) can be used to connect components in a 3D chip [33]. NoCs provide scalability when integrating multiple layers. NoCs also provide flexibility in terms of topologies. In this chapter, mesh, concentrated mesh and flattened butterfly topologies are explored as possible NoC topologies for 3D chips. In Section 4.2, TSVs are used to interconnect many-tiers of many-core chips. Various topologies are then investigated to realize performance benefits of each and also to compare the area occupied by TSVs. In Section 4.3, the results are concluded.

4.2 3D Network-on-Chip Analysis

All analysis presented in this chapter are for ITRS 2012 technology year [29]. However, the same analysis can be repeated for other technology years. Through silicon vias (TSVs) are assumed to interconnect the various tiers on the 3D chip. The properties of TSVs are extracted from ITRS. Figure 4.1 shows a detailed diagram of various TSV dimensions. As per ITRS 2012, the TSV diameter ($\emptyset_{TSV}$), contact pitch (P_{TSV}), bonding pad accuracy (Δ) and TSV keep out area are assumed to be 2 µm, 4 µm, 0.5 µm and 2 µm, respectively. Hence, the effective diameter of a TSV becomes 7 µm. TSVs are assumed to be 10 µm long [29]. For these lengths, the maximum achievable bandwidth for TSV is determined to be 11 THz [34]. However, this cannot be practically realized due to the driver switching limitation. Hence, for the rest of the chapter, each TSV is assumed to have a bandwidth equal to clock frequency of the chip.

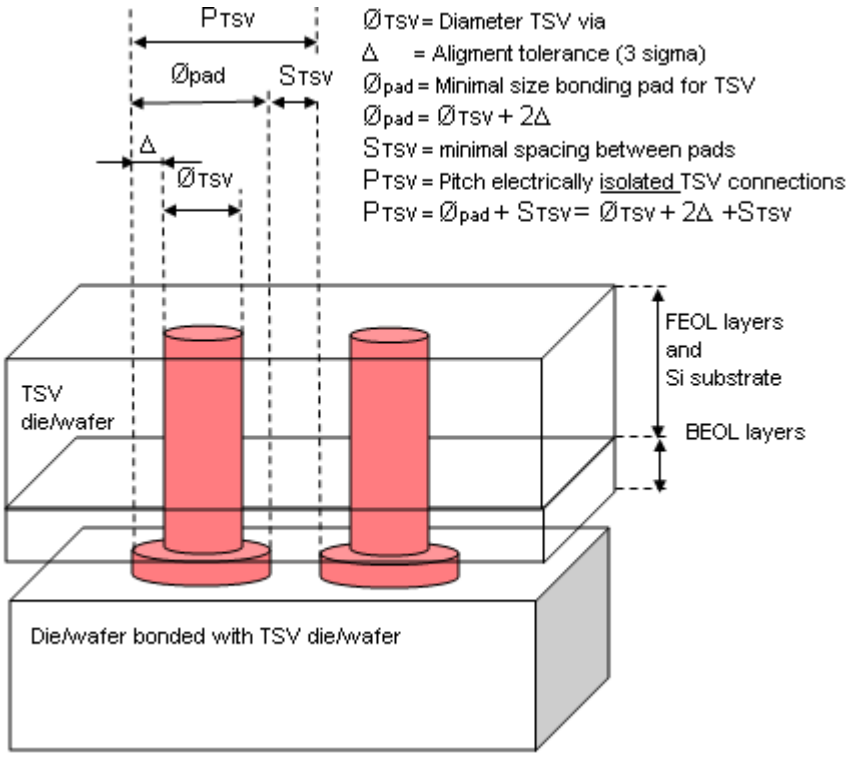


Figure 4.1: Dimensions of a TSV [29].

Face-to-back wafer bonding is preferred over face-to-back, to enable more than 2 tiers of 3D integration [35]. Figure 4.2 shows the difference between the two types of wafer bonding.

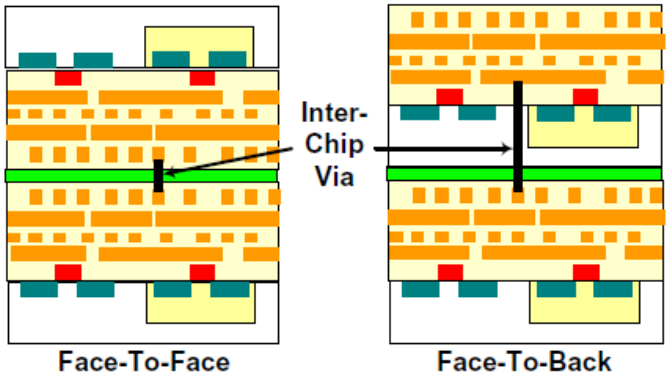


Figure 4.2: Face-to-face and face-to-back wafer bonding [35].

4.2.1 Mesh NoC Topology

A mesh topology is as shown in Figure 4.3. Initially, one tier is assumed (2D mesh). Later on, 3D mesh topologies are realized by 'folding' the existing 2D mesh to form additional tiers. For example, a 2 tier chip would be formed by folding the mesh topology in the middle. 4 tier chips are then realized by folding 2 tier chips, so on and so forth.

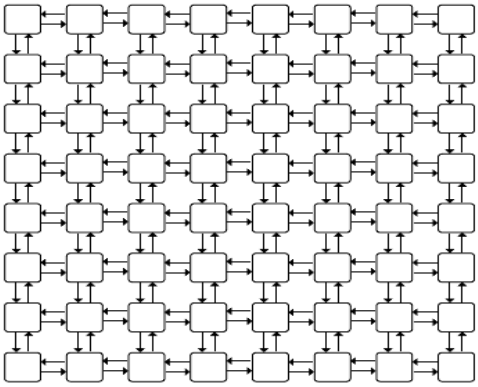


Figure 4.3: Mesh NoC topology [21].

The total time required to transport data from one corner of the chip to the diagonally opposite on the highest tier is defined as the worse case delay. It is assumed that all routers contribute a delay equal to 5 clock cycles and each horizontal or vertical interconnect has a delay of 1 clock cycle.

Since face-to-back wafer bonding is preferred in this chapter, it is important to analyze area lost to TSVs on the 2nd tier, as shown in Figure 4.2. Area occupied by TSVs are calculated by using the dimensions of TSVs that were determined earlier. The number of TSVs in the vertical dimension per core, is determined by the router area using ORION 2.0 [30]. The same analysis performed in Chapter 3 is repeated, by fixing the router area and then determining the maximum channel width. It is assumed that, the channel width is equal to the number of TSVs per core.

The worse case delay for many-core, many-tier mesh based chips are shown in Figure 4.4 for various number of tiers and cores per die.

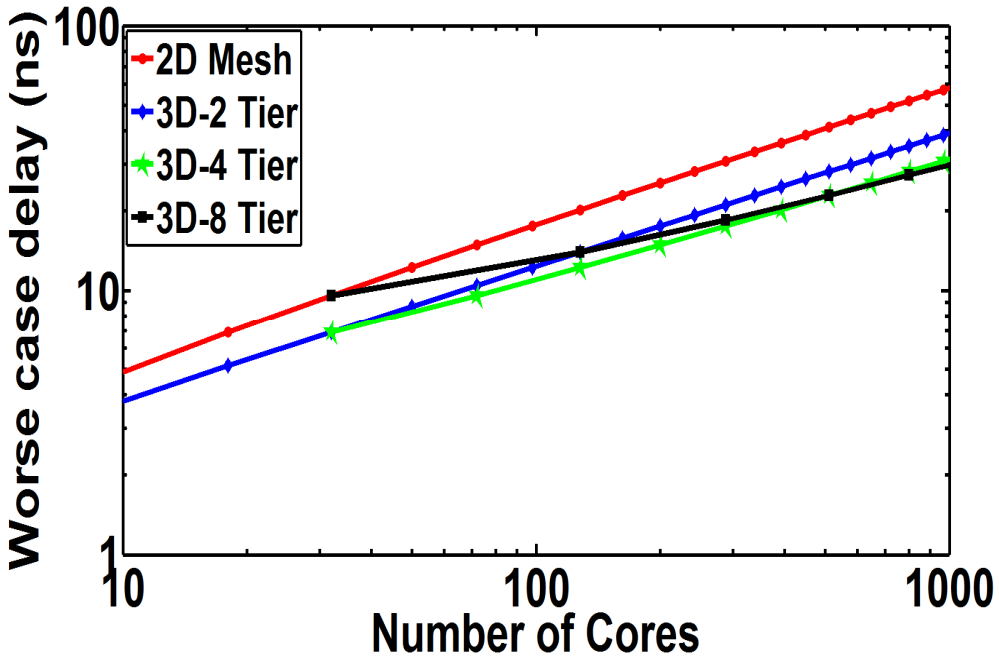


Figure 4.4: Worse case delay for mesh topology based many-core, many-tier chips versus number of cores for various number of tiers.

It can be observed from Figure 4.4 that, the advantage of reducing worse case delay by increasing the number of tiers from 2D mesh to a 2 tier chip is large. However, this improvement diminishes when the number of tiers are increased from 2 to 4 and then to 8. The active area lost, expressed as a fraction of one core area, to TSVs in the 2nd tier is shown Figure 4.5. The silicon area lost to TSVs in the 2nd tier is determined to be a very small fraction, less than 0.5%, of the core area.

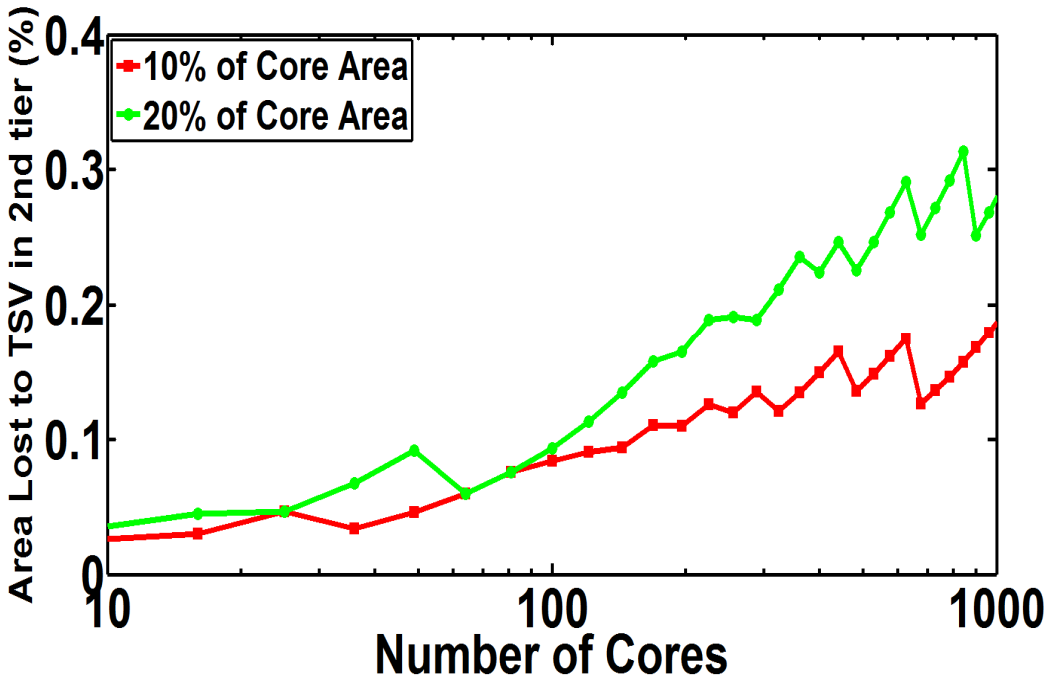


Figure 4.5: Active area lost to TSVs in the 2nd tier of a mesh topology based 3D chip, expressed as a fraction of one core area versus number of cores for 10% and 20% of router areas.

4.2.2 Concentrated Mesh NoC Topology

The analysis performed on mesh topology are repeated for concentrated mesh topology. Concentration factor of 4 is considered. However, the analysis can be extended for higher concentration factors. A concentrated mesh topology is shown in Figure 4.6.

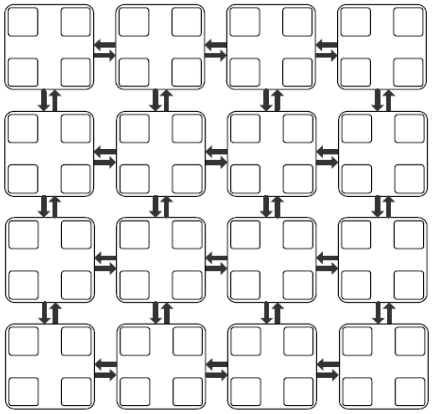


Figure 4.6: Concentrated mesh topology with a concentration factor of 4 [21].

The worse case delay for many-core, many-tier concentrated mesh based 3D chips are as shown in Figure 4.7.

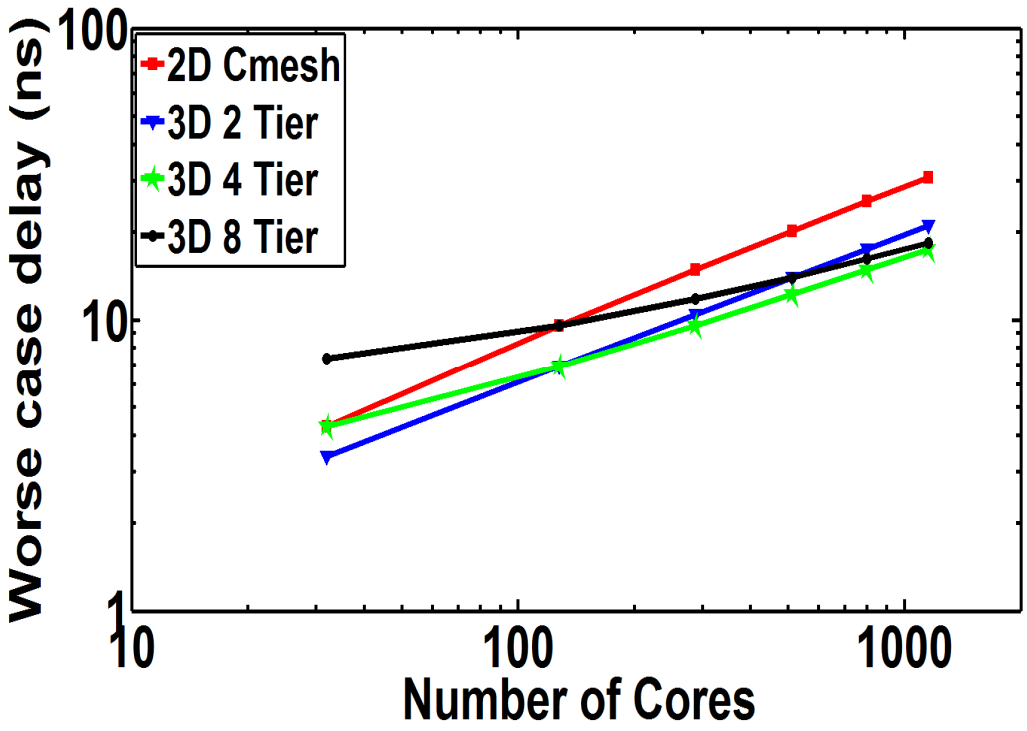


Figure 4.7: Worse case delay for concentrated mesh topology based many-core, many-tier chips versus number of cores for various number of tiers.

On comparing Figures 4.4 and 4.7 it can be seen that, concentration helps reduce worse case delay. This can be attributed for the reduction in number of hops in concentrated mesh topology for the same number of cores. However, it can also be observed that increasing the number of tiers, does not linearly reduce the worse case delay. In Figure 4.8, the area lost in the 2nd tier is shown for a concentrated mesh topology.

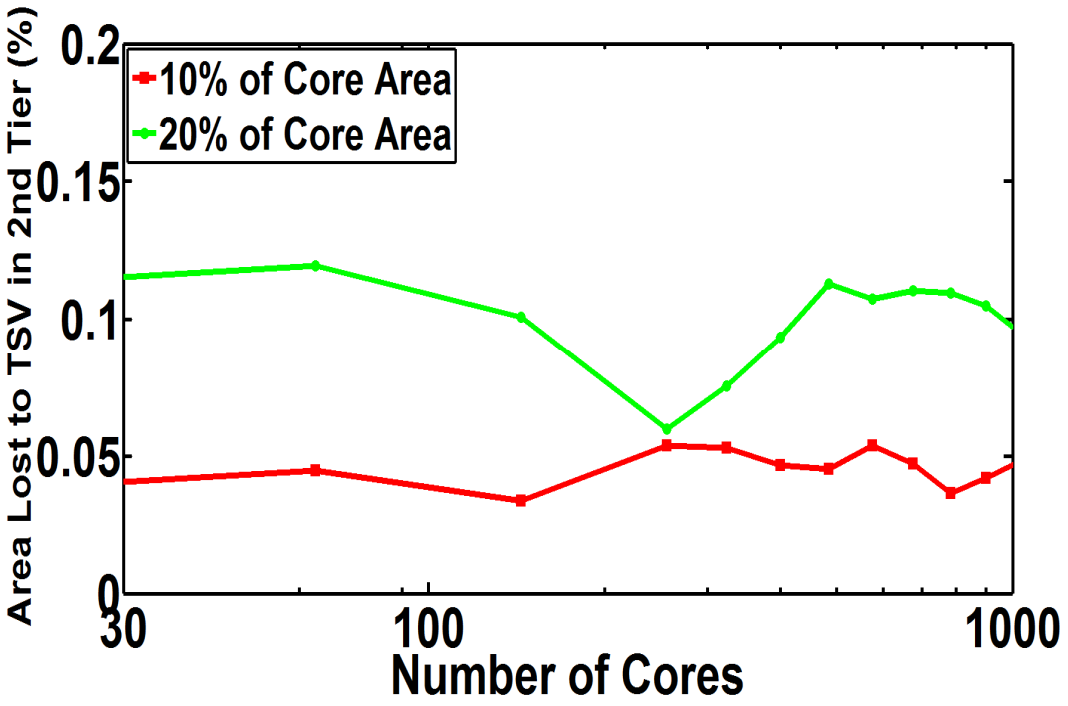


Figure 4.8: Active area lost to TSVs in the 2nd tier of a concentrated mesh topology based 3D chip, expressed as a fraction of one core area versus number of cores for 10% and 20% of router areas.

Active area lost to TSVs is again determined to be a very small fraction, less than 0.2%, of cores' area. In a concentrated mesh, a core refers to a grouping of 4 cores, as shown in Figure 4.8. Router area is found to be the bottleneck to increase number of TSVs per core.

4.2.3 Flattened Butterfly NoC Topology

A flattened butterfly based NoC topology is as shown in Figure 4.9.

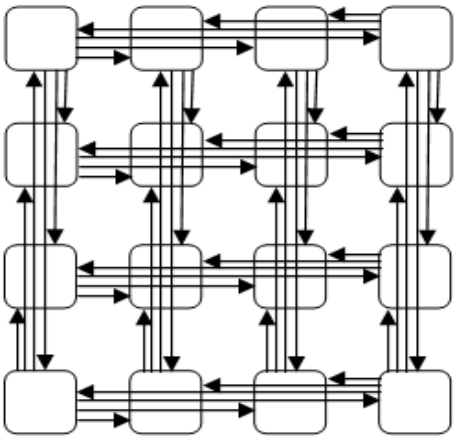


Figure 4.9: Flattened butterfly topology.

The worse case delay for a flattened butterfly topology will always remain the same: two hops through 3 routers, irrespective of number of tiers. This is because, a core in the lowest tier is directly connected to a core in the highest tier in the same dimension. However, it is worthwhile to have a look at the area that is lost to the TSVs in the 2nd tier. It can be predicted that, the active area lost would be more than that of mesh and concentrated mesh since the bottom most core is connected to every core on top of it. Figure 4.10 shows the area lost to TSV in the 2nd tier.

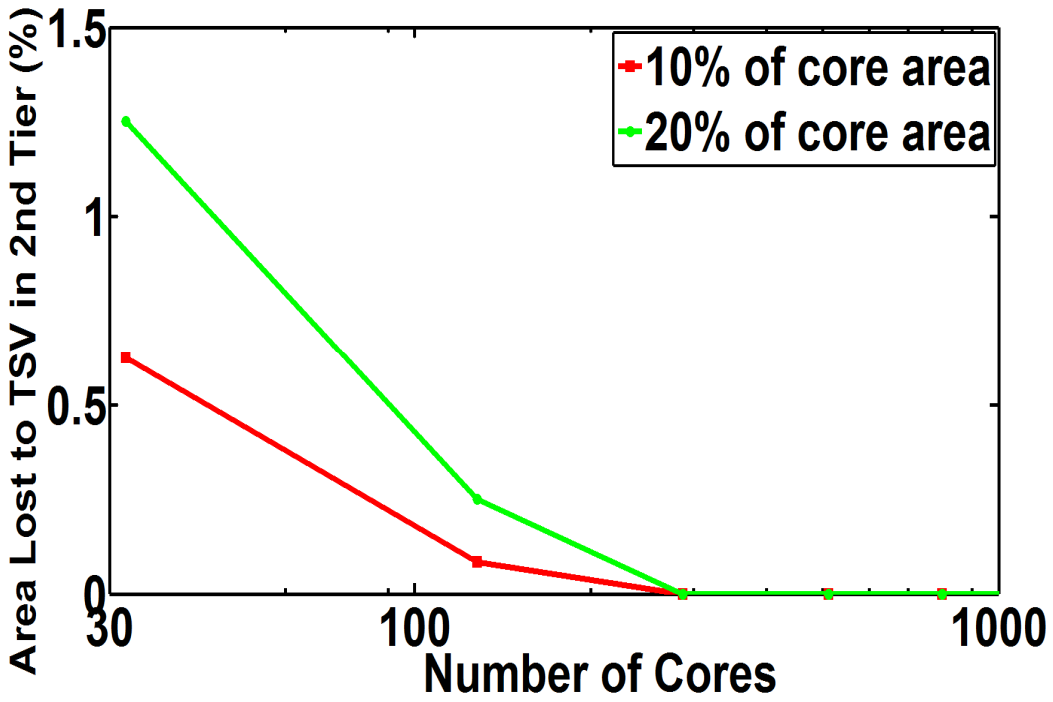


Figure 4.10: Active area lost to TSVs in the 2nd tier of a flattened butterfly topology based 3D chip, expressed as a fraction of one core area versus number of cores for 10% and 20% of router areas.

It can be observed that less than 2% of a cores' area is lost to TSVs. The number of TSVs go to zero for 256 cores or more, as the router area increases above 20% of core area. Once again, router area is found to be the limiting factor to increase the number of TSVs per core.

4.3 Conclusions

In this chapter, a study of different network-on-chip topologies for many-core many-tier 3D chips is performed. It is shown that, adding more tiers does not proportionately reduce the worse case delay for the same number of cores. Also, for the first time, the active area occupied by TSVs in the upper tiers has been quantified. It is concluded that router area is the bottleneck to increasing number of TSVs per core.

Chapter 5

Future Work and Conclusion

In this chapter, some of the potential extensions to this thesis that can be pursued in the future are presented. These extensions include hierarchical network-on-chip topology, wiring demand for network-on-chip routers, scaling of routers and wires and the study of 3D caches. At the end of this chapter, the key conclusions of this thesis are summarized.

5.1 Hierarchical Network-on-Chip Topologies

Mesh, concentrated mesh, flattened butterfly and concentrated flattened butterfly are the network-on-chip topologies studied in this thesis. However, for larger numbers of cores, a combination of various topologies might be required for scalability. Concentration factors could be as high as 16, with those 16 cores connected in a mesh topology. The concentration could then be interconnected by mesh, flattened butterfly or other topologies to form a network-on-chip. The analysis presented in this thesis can be easily extended to any other topology.

5.2 Wiring Demand for Network-on-Chip Routers

Throughout this thesis, it is determined that network-on-chip routers would be the crucial component for designing higher performance processors. Therefore, it is imperative to look into the building blocks of a NoC router. The wiring demand analysis needs to be extended for NoC routers since crossbars within routers are found to be wiring intensive. The size of the crossbar would then depend on the available wiring area for routers and hence could possibly limit the number of ports per router.

5.3 Scaling of Routers and Wires

It would be interesting to look at how routers and wires would scale with every technology year. Area and power dissipation of routers would be the focus to determine the right router architecture. Also, with the post CMOS era fast approaching, it would be imperative to study the characteristics of routers and wires made of novel materials such as, carbon nanotubes, graphene nano ribbons or silicon nanowires, to name a few.

5.4 3D Caches

One of the advantages of 3D integration is that it brings the memory closer to the processor. Possible 3D architectures that have been discussed include, a layer of CPU cores with multiple layers of memory stacked on top. Since TSVs have a very large bandwidth, data from memory could be brought to the processor quicker than existing L1 caches. Therefore, to know the advantages and limitations of memory stacking, it is worthwhile to model cache bandwidth requirements for many-core chips and then determine the feasibility of 3D caches.

5.5 Conclusion of Thesis

The main objectives of this thesis are: 1) to optimize global interconnects for many-core chips, 2) compare different network-on-chip topologies from a technological perspective and 3) to determine wiring demand for 2D and 3D NoC based chips. The main contributions of this thesis are as follows:

1. A circuit-aware interconnect technology optimization is performed for mesh based network-on-chip in many-core architecture. Optimal wire-width, W_{opt}, for a 144- core chip in the technology year 2015 is found to be more than 10 times smaller than previous optimization results where router latency, reduced

interconnect length and size effects of copper wires were ignored. The optimal wire width does not vary significantly with technology year, but is a strong function of number of cores on a die and the frequency of operation.

2. Optimization of global interconnects is then performed for a NoC in many-core architecture, after imposing practical limitations, to minimize energy-per-bit and delay and maximize bandwidth density, simultaneously. For a 1000 core chip in 2015, the optimal dimensions are minimum pitch and minimum width limited. The optimal number of repeaters is around 0.2 times the optimal number of repeaters found based on intrinsic delay of interconnects. This along with width and spacing optimization achieves a crucial 16% reduction in energy per bit.
3. A technology-cum-architecture aware network-on-chip interconnect optimization is performed for many-core chips. Optimal wire dimensions are determined for mesh, concentrated mesh, flattened butterfly and concentrated flattened butterfly topologies. Then using ORION 2.0, router area and router energy-per-bit are deduced. It is shown that router area is the main bottleneck to achieve larger core-to-core bandwidth. Flattened butterfly and concentrated flattened butterfly topologies have the best wiring area utilization, maximum bisectional bandwidth and the lowest worse case delay. However, these advantages diminish for larger number of cores due to the rapid increase in router area. Mesh and concentrated mesh topologies are found to be more scalable in terms of number of cores than flattened butterfly and concentrated flattened butterfly topologies.
4. A study of different network-on-chip topologies for many-core many-tier 3D chips is performed. It is shown that, adding more tiers does not proportionately reduce the worse case delay for the same number of cores. Also, for the first time, the active area occupied by TSVs in the upper tiers has been quantified. It is concluded that router area is the bottleneck to increasing number of TSVs per core.

List of Publications

[1] A. Balakrishnan and A. Naeemi, "Optimal Global Interconnects for Networks-on-Chip in Many-Core Architectures," *Semiconductor Research Corporation TECHCON*, Sep. 2009.

[2] A. Balakrishnan and A. Naeemi, "Optimal Global Interconnects for Networks-on-Chip in Many-Core Architectures," *IEEE Electron Device Lett*., vol. 31, no. 4, pp. 290–292, Apr. 2010.

[3] A. Balakrishnan and A. Naeemi, "Bandwidth, Delay and Energy aware Optimization of Global Interconnects for Many-Core Architectures," *IEEE Int. Interconnect Technology Conf*., Jun. 2010.

[4] A. Balakrishnan and A. Naeemi, "Bandwidth, Delay and Energy aware Optimization of Global Interconnects for Many-Core Architectures," *SemiconductorResearch Corporation TECHCON*, Sep. 2010.

[5] A. Balakrishnan and A. Naeemi, "Interconnect Network Analysis of Many-Core Chips," *IEEE Trans. Electron Dev*., in progress.

[6] A. Naeemi and A. Balakrishnan, "Interconnect Networks in 2D and 3D Nanoelectronic Systems," *IEEE/ACM Workshop on Variability Modeling and Characterization*, Nov. 2010, (invited talk).

References

[1] G.E. Moore, "Progress in digital integrated circuits," *IEEE IEDM Tech. Digst*., Dec. 1975, pp. 11-13.

[2] Intel Corporation. Available: http://www.intel.com.

[3] Intel Developer Forum, 2004.

[4] S. Borkar, "Thousand Core Chips: A Technology Perspective", *in Proc. Design Automation Conf*., pp. 746-749, 2007.

[5] Dell Inc. Available: http://www.dell.com.

[6] J. Bautista, "Tera-scale Computing and Interconnect Challenges", *in Proc. Design Automation Conf*., pp. 665-667, 2008.

[7] W. J. Dally, Keynote Speech, *in Proc. Design Automation Conf*., 2009. Available: http://videos.dac.com/46th/wedkey/dally.html.

[8] A. Naeemi, R. Venkatesan and J. D. Meindl, "Optimal Global Interconnects for GSI", *IEEE Trans. Electron Devices*, vol. 50, no. 4, pp. 980-987, Apr. 2003.

[9] X. C. Li, J. F. Mao, H. F. Huang and Y. Lui, "Global interconnect width and spacing optimization for latency, bandwidth and power dissipation," *in IEEE Electron Device Meeting*, pp. 2272-2279, 2005.

[10] H. Cho, K. H. Koo, P. Kapur, and K. C. Saraswat, "The Delay, Energy, and Bandwidth Comparisons between Copper, Carbon Nanotube, and Optical Interconnects for Local and Global Wiring Application," *in Proc. Int. Interconnect Technol. Conf*., pp. 135-137, Jun. 2007.

[11] H. Cho, K. H. Koo, P. Kapur and K. C. Saraswat, "Modeling of the Performance of Carbon Nanotube Bundle, Cu/Low-K and Optical On-chip Global Interconnects," *in Proc. Syst. Level Interconnect Prediction*, pp. 81-88, Mar. 2007.

[12] H. Cho, K. H. Koo, P. Kapur, and K. C. Saraswat, "Performance Comparisons Between Cu/Low-K Carbon-Nanotube, And Optics for Future On-Chip Interconnects," *IEEE Electron Device Lett*., vol. 29, no. 1, pp. 122-124, Jan. 2008.

[13] A. F. Mayadas and M. Shatzkes, "Electrical-Resistivity Model for Polycrystalline Films: The case of arbitrary reflection at external surfaces," *Phys. Rev. B*, vol. 1, no. 4, pp. 1382-1389, Feb. 1970.

[14] E. H. Sondheimer, "Mean Free Path of Electrons in Metals," *Advances in Physics*, vol. 50, no. 6, pp. 499-537, Sep. 2001.

[15] Steinhogl. W, Schindler. G, Steinlesberger. G, Traving. M, Enelhardt. M, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100nm and smaller," *J. Appl. Phys*., vol. 97, no. 2, pp. -, 2005.

[16] R. Venkatesan, J. A. Davis, and J. D. Meindl, "Compact Distributed RLC Interconnect Models—Part IV: Unified Models for Time Delay, Crosstalk, and Repeater Insertion," *IEEE Trans. Electron Devices*, vol. 50, no. 4, pp. 1094-1102, Apr. 2003.

[17] International Technology Roadmap for Semiconductors (ITRS), 2008. Available: http://www.itrs.net/

[18] M. J. Kobrinsky, B. A. Block, J. F. Zheng, B. C. Barnett, E. Mohammed, M. Reshotko, F. Robertson, S. List, I. Young and K. Cadien, "On-Chip Optical Interconnects," *Intel Technol. J*., vol. 8, no. 2, pp. 129-142, May 2004.

[19] W. J. Dally and B. Towels, "Route Packets, Not Wires: On-Chip Interconnection Networks," *in Proc. Design Automation Conf*., pp. 684-689, 2001.

[20] Intel Tera-Scale Research.
Available: http://techresearch.intel.com/articles/Tera-Scale/1489.htm.

[21] B. Grot and S. W. Keckler, "Scalable On-Chip Interconnect Topologies," *Chip Multiprocessor Memory Systems and Interconnects*, Jun. 2008.

[22] H. Elmiligi, M. W. El-Kharashi and F. Gebali, "A Delay Model for Networks-on-Chip Output Queuing Router," *in Proc. IEEE 6th International Workshop on System on Chip for Real Time Applications*, pp. 95-98, Dec. 2006.

[23] L. S. Peh and W. J. Dally, "A Delay Model and Speculative architecture for Pipelined Routers," *in Proc. of the 7th International Symposium on High-Performance Computer Architecture*, pp. 255-266, 2001.

[24] R. Das, S. Eachempati, A. K. Mishra, V. Narayanan, C. R. Das, "Design and Evaluation of a Hierarchical On-Chip Interconnect for Next-Generation CMPs," *in Proc. High Performance Computer Architecture*, pp. 175-186, 2009.

[25] R. Sarvari, "Impact of Size Effects and Anomalous Skin Effect on Metallic Wires as GSI Interconnects," *PhD. Dissertation*, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 2008.

[26] T. Sakurai and K. Tamaru, "Simple Formulas for Two- and Three-Dimensional Capacitances," *IEEE Trans. Electron Devices*, vol. 30, no. 2, pp. 183-185, Feb. 1983.

[27] R. Venkatesan, J. A. Davis, K. A. Bowman, J. D. Meindl, "Optimal n-tier Multilevel Interconnect Architectures for Gigascale Integration (GSI)," *IEEE Tran. VLSI Sys*., vol. 9, no. 6, December 2001.

[28] A. Balakrishnan and A. Naeemi, "Optimal Global Interconnects for Networks-on-Chip in Many-Core Architectures", *IEEE Electron Device Lett*., vol. 31, no. 4, pp. 290–292, Apr. 2010.

[29] International Technology Roadmap for Semiconductors (ITRS), 2009. Available: http://www.itrs.net/

[30] A. B. Kahng, B. Li, L. S. Peh and K. Samadi, "ORION 2.0: A Fast and Accurate NoC Power and Area Model for Early-Stage Design Space Exploration," *in Proc. Design Automation and Test*, pp. 423-428, Apr. 2009.

[31] K. Sakuma, P. S. Andry, C. K. Tsang, S. L. Wright, B. Dang, C. S. Patel, B. C. Webb, J. Maria, E. J. Sprogis, S. K. Kang, R. J. Polastre, R. R. Horton, J. U. Knickerbocker, "3D chip-stacking technology with through-silicon vias and low-volume leadfree interconnections," *IBM J. Res. & Dev*., Vol. 52, No. 6, pp. 611-622, Nov. 2008.

[32] P. G. Emma, E. Kursun, "Is 3D chip technology the next growth engine for performance improvement?," *IBM J. Res. & Dev*., Vol. 52, No. 6, pp. 541-552, Nov. 2008.

[33] S. Murali, L. Benini and G. De Micheli, "Design of Networks on Chips for 3D ICs," *in Proc. Design Automation Conf*., pp. 167-168, 2010.

[34] D. E. Khalil, Y. Ismail, M. Khellah, T. Karnik, and V. De, "Analytical Model for the Propagation Delay of Through Silicon Vias," *in Proc. Int. Symp. on Quality Elec. Des*., pp. 553-556, 2008.

[35] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan and M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," *in Proc. Int. Symp. on Comp. Arch*., 2006.